Open Research Online

The Open University's repository of research publications
and other research outputs

Statistical aspects of credit scoring

Thesis

How to cite:

Henley, William Edward (1995). Statistical aspects of credit scoring. PhD thesis The Open University.

For guidance on citations see FAQs.

© 1994 The Author

https://creativecommons.org/licenses/by-nc-nd/4.0/

Version: Version of Record

Link(s) to article on publisher's website:
http://dx.doi.org/doi:10.21954/ou.ro.0000e061

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online's data policy on reuse of materials please consult the policies page.

oro.open.ac.uk

DX184766
UNRESTRICTED

Statistical aspects of credit scoring

William Edward Henley, BA[1]

Department of Statistics,
Faculty of Mathematics and Computing,
The Open University.

2nd December 1994

Author number : M7110877
Date of submission: 8 December 1994
Date of award : 22 March 1995

---

[1]A thesis submitted for the degree of Doctor of Philosophy

Contents

Part 1: An introduction to credit scoring

Part 3: Approaches to classifier design

Abstract

This thesis is concerned with statistical aspects of credit scoring, the process of determining how likely an applicant for credit is to default with repayments. In Chapters 1-4 a detailed introduction to credit scoring methodology is presented, including evaluation of previous published work on credit scoring and a review of discrimination and classification techniques.

In Chapter 5 we describe different approaches to measuring the absolute and relative performance of credit scoring models. Two significance tests are proposed for comparing the bad rate amongst the accepts (or the error rate) from two classifiers.

In Chapter 6 we consider different approaches to reject inference, the procedure of allocating class membership probabilities to the rejects. One reason for needing reject inference is to reduce the sample selection bias that results from using a sample consisting only of accepted applicants to build new scorecards. We show that the characteristic vectors for the rejects *do not contain information about* the parameters of the observed data likelihood, unless extra information or assumptions are included. Methods of reject inference which incorporate additional information are proposed.

In Chapter 7 we make comparisons of a range of different parametric and non-parametric classification techniques for credit scoring: linear regression, logistic regression, projection pursuit regression, Poisson regression, decision trees and decision graphs. We conclude that classifier performance is fairly insensitive to the particular technique adopted.

In Chapter 8 we describe the application of the k-NN method to credit scoring. We propose using an adjusted version of the Euclidean distance metric, which is designed to incorporate knowledge of class separation contained in the data. We evaluate properties of the k-NN classifier through empirical studies and make comparisons with existing techniques.

Acknowledgements

I would like to thank my supervisor, Professor David Hand, for all his help, encouragement and valuable advice during the preparation of this thesis. I would also like to thank Ivor Langley and Mark Goodchild from the Littlewoods Organisation for acting as my industrial supervisors and providing me with a practical perspective on credit scoring. I am indebted to Alan Rogers, Helen Lawson, John Gutherie and others who have helped with data preparation during my regular visits to the Littlewoods head office in Liverpool.

I would like to thank all the people at the Open University who have helped me during my three year studentship, especially Fergus Daly and Jon Oliver for passing on some of their computing expertise.

I am very grateful to my family and friends, in particular to my parents David and Liz Henley, my brother Anthony and my wife Rachel, for all their support over the last three years.

PART 1

An introduction to credit scoring

Chapter 1

Introduction

Consumer credit is granted by banks, building societies, retailers, mail order companies and various other lending institutions and is a sector of the economy that has seen rapid growth over the last thirty years. Traditional methods of credit risk assessment involved the use of human judgement, based upon experience of previous decisions, to determine whether to grant credit to a particular individual. The economic pressures resulting from the increased demand for credit and the emergence of new computer technology has led to the development of sophisticated statistical models to aid the credit granting decision. *Credit scoring* is the name used to describe the process of determining how likely an applicant is to default with repayments. Statistical models which give estimates of these default probabilities are referred to as *scorecards* or *classifiers*. Standard methods used for developing scorecards are discriminant analysis, linear regression, logistic regression and decision trees. An accept/reject decision can then be taken on a particular applicant by comparing the estimated good/bad probability with a suitable threshold.

Despite the widespread use of credit scoring techniques in the consumer credit industry, there are several aspects of the methodology that have not received sufficient attention in the literature. The two main reasons for this are commercial confidentiality and the lack of widely available data sets. In this thesis we examine statistical aspects of credit scoring with the aid of a real data set from a large mail order company[2]. Our attention is focused on methodological and conceptual issues of relevance to the general credit granting decision, rather than the prescription of models for specific populations or data sets.

In Part 1 of this thesis we provide a detailed introduction to credit scoring methodology:

---

[2] See Henley and Hand (1995) for an overview of statistical credit scoring

- In Chapter 2 we provide an outline of the consumer credit granting process. This includes a discussion of the objectives of credit scoring and the nature of credit data. We also describe the model construction process from a practical perspective.

- In Chapter 3 we review previous published work on credit scoring and identify areas needing more extensive research.

- In Chapter 4 we review discrimination and classification techniques and assess their suitability for constructing credit scoring models.

The areas of research described in this thesis fall into two categories.

First, in Part 2 we consider fundamental aspects of credit scoring methodology:

(1) *Assessment of classifiers.*
Before we can draw conclusions about the suitability of different techniques for credit scoring, we need some criterion for performance. Although we concentrate on a specific criterion in this thesis (the minimisation of bad rate for a particular acceptance rate), we include discussion of general performance criteria. In particular we distinguish between two ways of assessing the performance of a classifier: absolute and relative performance. In Chapter 5 we propose two tests for comparing classifiers: a likelihood ratio test and a significance test which uses Fisher's Exact test. The two tests address subtly different questions and can be seen as complimentary. Both tests can also be adapted to assess different performance criteria, such as bad rate amongst the accepts and error rate.

(2) *Reject Inference.*
Reject Inference is the name for the procedure of allocating probabilities of becoming a good or a bad to the rejects. One reason for wanting to do this is to help reduce the sample selection bias that results from using a sample consisting only of accepted applicants to build new scorecards. In Chapter 6 we examine methods of reject inference that have been proposed in the literature. Much of the work seems to have been based upon a poor understanding of what can be achieved using the rejects. We aim to clarify this

and propose some new methods which require additional information or assumptions.

Secondly, in Part 3 we assess the merits of different approaches to classifier design:

(1) *A comparison of different classification techniques for credit scoring.*
In Chapter 7 we discuss the application of linear and logistic regression to credit scoring. Both of these techniques are widely used by developers of credit scoring systems. However, as we discuss, logistic regression might be considered more appropriate for credit scoring data than linear regression. After considering the theoretical merits of the two methods we present an empirical comparison. Further comparisons are made with a range of parametric and non-parametric classification techniques including decision trees, projection pursuit regression and Poisson regression. Classifier performance is found to be fairly insensitive to the particular technique used.

In Chapter 7 we also explore how the relative performance of the classification techniques change as the definition of credit default changes. This issue is addressed by considering the construction of classifiers to identify fraudulent applications for credit. The area of fraud is one that is becoming of increasing importance to credit grantors. Comparisons are made between fraud classifiers and standard good/bad classifiers.

(2) *The proposal of a new approach to credit scoring: the k-Nearest Neighbour method.*
The *k*-Nearest Neighbour (*k*-NN) method is a standard technique in pattern recognition and non-parametric statistics. In Chapter 8 we consider reasons why it may be a suitable technique for building a credit scoring model. Part of the implementation of the *k*-NN method is the selection of a suitable distance measure. We propose using an adjusted version of the Euclidean distance metric which attempts to incorporate knowledge of class separation contained in the data. We evaluate properties of the *k*-NN classifier through empirical studies and make comparisons with a range of standard credit scoring techniques.

Chapter 2

An outline of the consumer credit granting process

2.1 The objectives of credit scoring

In this thesis the term "credit" is used to refer to an amount of money that is loaned to a consumer by a financial institution, for a fixed period of time. It is assumed that the consumer is required to pay off a proportion of the loan each month.

We are primarily concerned with the construction of credit scoring models to estimate the probability of an applicant for credit defaulting with repayments (this process is sometimes called "application scoring"). The model takes the form of a relationship between creditworthiness and a number of predictor variables, extracted from the application form and other sources. These variables are called *characteristics* and the values they take are referred to as *attributes*. The model or scorecard is constructed using a sample of applicants, for each of whom the true creditworthiness is known.

We introduce some notation which will be used throughout the thesis. For a particular applicant with characteristic vector $\mathbf{x}$, a credit scoring model gives an estimate of the predicted probability of that applicant being a good risk, given by $\hat{P}(g|\mathbf{x})$. An accept/reject decision is then taken by comparing $\hat{P}(g|\mathbf{x})$ with a threshold.

The usual objective of credit scoring is to produce a model which provides the best possible discrimination between good and bad applicants. Various methods of measuring the discriminability of a scorecard are possible. In particular, the criterion adopted in this thesis is the minimisation of the bad rate amongst accepted applicants for credit (see Chapter 5 for more discussion of this criterion). Other criteria, such as the minimisation of the error rate of the classifier, could also be adopted.

The objective considered above includes the assumption that applicants for credit can be split into two distinct classes: those with a high probability of defaulting (the "bads") and those with a low probability (the "goods"). This allows us to treat the prediction of creditworthiness as a two state problem and to exploit techniques for dealing with binary response data. It might be more appropriate to consider creditworthiness as a continuous property or to distinguish between different types of defaulter. The effect of alternative definitions of credit default is considered by Crook et al. (1992). In Section 7.4 we consider the construction of scorecards to predict fraud using a tighter definition of credit default. Further research is needed into this aspect of credit scoring to determine the implications of different representations of credit behaviour.

Another standard assumption used in credit scoring is that the true class of each credit applicant is fixed. In practice, the credit repayment behaviour of an applicant may vary according to factors, such as employment status, which can change over time. Models which allow a change of state between the good and bad classes are considered in Section 4.12.

It can be argued that minimisation of bad debt is only one dimension of the credit granting decision. Credit scoring techniques can be used to address more general business objectives. For example:

- The main aim of the lender is usually to maximise profits. This objective can conflict with minimising bad debt, because some high risk applicants can be very profitable. Credit scoring methodology can be used to provide models to estimate the profitability of an applicant for credit. This information can be used to make an accept/reject decision. Alternatively, it could be combined with an estimate of the probability of an applicant defaulting, as part of a more complex decision making strategy. One potential difficulty with "profit scoring" is the specification of a suitable definition of the profitability of an individual applicant.

- Credit scoring techniques can be used to build up profiles of existing customers and to make decisions about whether to offer new products. This is referred to as "behavioural scoring".

Application scoring is the most common form of credit scoring and we restrict attention to it in this thesis because of the many unresolved methodological issues relating to it. However, much of the discussion applies equally to profit and behaviour scoring. A general review of literature on the objectives of credit scoring is presented in Section 3.2.

2.2 Description of credit scoring data

The data used in our analysis consists of samples from the full population of applicants for credit from a large mail-order company. Several different samples are used in this thesis in order to assess different aspects of credit scoring. For example, in Chapter 8 we use samples from two different time periods in order to assess the robustness of our proposed k-Nearest Neighbour method to changes in the population. Table 2.1 shows a description of the standard sample used.

	Number of variables	Number of classes	Number of cases	% of bads in full sample
Design set	65	3	15728	31.01
Test set	65	3	4132	31.34

Table 2.1: A description of a standard credit scoring data set.

The full sample was randomly split into a design set (80%) and a test set (20%). The 80/20 holdout split was adopted to allow consistent comparisons with the credit grantor's existing results. As discussed in Section 3.1.1, Reichert et al. (1983) demonstrate that the holdout procedure gives stable mean error rates over a range of holdout percentages (including 20%). Furthermore, our approach can be justified because of the large numbers of applicants in both the design and holdout samples, thus enabling accurate measurement of performance. We discuss assessment of performance in Section 2.4 and Chapter 5.

For each sample, the following procedure was used to define creditworthiness. The entire sample, including applicants who would normally have been rejected, were given credit and observed over a period of a year. Applicants who defaulted for three consecutive months were classified as bad and the remaining applicants in our dataset were either classified as good or to an intermediate class (if one or

two payments were outstanding at the end of the observation period). The intermediate class was discarded from our analysis to make our results useful to the credit grantor and to satisfy commercial objectives. (Discarding "other" classes is a common practice in the credit industry; see, for example, Crook et al. (1992)). We put forward our own interpretation of the commercial motives for omitting the intermediate class:

- The credit grantor is primarily interested in the relative proportions of good and bad risks accepted. The business depends upon balancing the conflicting aims of maximising sales to good customers and minimising bad debt.

- The true status of the intermediate class is unclear and, thus, the credit grantor has ambivalent feelings about whether to accept or reject them. Removing them from the analysis allows one to concentrate on discrimination between the good and bad applicants.

Although the above approach is not ideal, it was the problem we were asked to solve. Moreover, there is not an obvious way to include three classes (goods/bads/others) in the analysis when our aim is to split future applicants into two classes (accepts/rejects). This problem is somewhat analagous to that of reject inference (see Chapter 6). There is a need for further research into the whole area of definitions of creditworthiness and into the specific issue of how to treat the "other" classes when creditworthiness is defined as above. More complex credit granting strategies could attach some relative importance to the others (or even measure creditworthiness on some continuous scale) in order to include them in the model construction.

In practice, rejected applicants are not given credit and so their true status cannot be determined. This means that a credit grantor only has a sample of accepted applicants with which to construct new scorecards. A new scorecard based solely on the accepted applicants may be biased. A standard approach to this problem is to infer the status of the rejects and combine this information with the known status of the rejects to reduce the bias in the new scorecard. "Reject Inference" is the name given to the process of attempting to infer the $\hat{P}(g|\mathbf{x})$ for the rejects. In Chapter 6 we consider methods of reject inference that have been proposed and examine the theoretical assumptions that underpin them. In particular we show that the characteristic vectors for the rejects do not contain information

about the parameters of a credit scoring model unless additional information or assumptions are included. Because we have access to a sample with the true status known for the rejects, we are able to to ignore the question of reject inference when assessing classifier performance in other chapters.

Characteristics for inclusion in credit scoring models are extracted from the application form and databases containing information on the credit history of households throughout the country. Table 2.1 shows that 65 characteristics were available for use in our analysis (these characteristics are not all available at the initial vetting stage described in the next section). Table 2.2 presents a range of typical characteristics available for scorecard development. For reasons of commercial confidentiality these characteristics do not correspond exactly to those used in our analysis.

Characteristic	Title	Attributes
1	Time at present address	0-1, 1-2, 3-4, 5+ years
2	Home status	Owner, tenant, other
3	Postcode	Band A, B, C, D, E
4	Telephone	Yes, no
5	Marital status	Single, married, divorced
6	Applicant's annual income	£(0-10k), £(11-20k), £21k+
7	Credit card	Yes, no
8	Type of bank account	Cheque and/or savings, none
9	Age	18-25, 26-40, 41-55, 55+

Table 2.2: Description of credit scoring characteristics.

The characteristics available for inclusion in scorecards are usually nominal or ordinal in nature. (Suitable examples from Table 2.2 are "Marital Status" and "Time at address" respectively.) Characteristics which are nominal need to undergo some form of processing before they can be used as predictor variables for standard credit scoring techniques. Several ways of doing this have been proposed:

(1) Indicator (dummy) variables can be used to represent each attribute of a nominal characteristic. For a characteristic with r categories this involves the introduction of $(r - 1)$ binary variables. As r becomes large there is a danger of

overfitting the data. Another weakness of this approach for methods that require the predictor variables to be multivariate normal (such as linear discriminant analysis), is that indicator variables violate this assumption.

(2) Krzanowski (1975) advocates building a different discriminant function for each possible combination of the nominal variables. As the number of nominal variables becomes large this approach becomes impractical.

(3) Crook et al. (1992) and Boyle et al. (1992) describe approaches to transforming a qualitative variable into a quantitative one. We briefly describe the proposed transformations.

For a nominal characteristic with r attributes let g_i be the number of goods in attribute i and let b_i be the corresponding number of bads. Let the total number of goods and bads in the sample be given by:

$$G = \sum_{i=1}^{r} g_i \quad \text{and} \quad B = \sum_{i=1}^{r} b_i$$

Then quantitative representations of the i th attribute are given by:

$$t_i = \begin{cases} g_i / b_i \\ g_i / (g_i + b_i) \\ g_i B / b_i G \\ \log(g_i B / b_i G) \\ \log(g_i / (g_i + b_i)) \end{cases} .$$

Methods (1) and (2) from above are subject to limitations when the number of attributes is large. For this reason we choose to adopt method (3) for dealing with nominal characteristics.

The particular transformation that we choose (the 4th in the above list for t_i) is widely used in the credit industry. We express it in the following form in this thesis. The values of the predictor variables are replaced by *weights of evidence* where the weight of evidence of the jth attribute of the ith characteristic is given by:

$$w_{ij} = \ln(p_{ij} / q_{ij})$$

where p_{ij} is the number of goods in attribute j of characteristic i divided by the total number of goods and q_{ij} is the number of bads in attribute j of characteristic i divided by the total number of bads.

One added advantage of using weights of evidence to represent nominal characteristics is that it orders the attributes so that they have a monotone relationship with creditworthiness in the sample. Further work is needed to identify a transformation giving an optimal ordering and spacing of the attribute values.

For the data sets used in this thesis, all characteristics, including continuous ones, were categorised and put into weights of evidence form. This allows us to treat all characteristics in the same way and exploit standard classification methods. This simplification of the data is permissible because we are interested in prediction rather than description of the data.

A decision tree combining several of the other available characteristics was included in the analysis as a characteristic. This was included because the credit grantor in our problem uses it and we wanted to make things as useful as possible to them. From a theoretical point of view this helps models that fit a linear combination of the predictors (such as linear and logistic regression) to take account of interactions between variables. This approach is similar to the hybrid classifier of discriminant analysis and decision trees proposed by Boyle et al. (1992).

There are two other general properties of credit scoring data: high correlations between variables (we return to this issue in Section 2.3.1) and missing values in the predictor variables. Methods for dealing with missing data, such as the EM algorithm can be used to alleviate this second problem. (The EM algorithm is an algorithm for finding maximum-likelihood solutions when there are missing data. See Dempster et al. (1977) for more details). In this thesis we take a different approach to dealing with missing characteristic values: an attribute "no information" is included as an additional attribute for characteristics with some missing values. This is a suitable approach as long as the proportion of the sample with missing values is not too large. One reason for treating missing values in this way is that the mere fact that an item is missing can be predictive.

2.3 Construction of credit scoring models

In this section we consider the construction of a credit scoring model using linear regression. We give a brief description of some currently used approaches to credit scoring development. The credit grantor, in our specific problem, uses a two stage process for assessing creditworthiness. Credit applicants are assessed using an initial vetting system (a "mini" scorecard). This results in three possible outcomes: accept outright, reject outright or request further information. The third option involves applicants filling in a more detailed application form and being scored up under a "full" scorecard. This two stage approach is taken for practical commercial reasons. In order to concentrate on methodological issues, we restrict attention to the initial stage of vetting throughout this thesis. We return to the idea of a multi-stage vetting system in Section 8.2.5.1 where we discuss the reject option for the k-NN method.

2.3.1 Variable selection

There are two main dangers of using highly correlated characteristics in a scorecard: first, it may be difficult to determine the magnitude of the effect that each individual characteristic has on creditworthiness. This is important as the scorecard developer may be required to justify the contribution to the model of each characteristic on business grounds. Secondly, it may lead to "inversions" in the resulting scores; this means that for a particular characteristic expected score rankings may be reversed (for example, giving more weight to an applicant who does not have a current bank account than to one who does). These factors provide a motivation for reducing the number of characteristics to an efficient level and for finding ways of identifying the most predictive subsets.

Furthermore, as the number of characteristics becomes very large there is a danger that the classifier will overfit the training data, particularly when the sample size is small. (This consideration is not very important in our problem because the sample sizes are relatively high). It is also desirable to limit the number of variables in order to reduce application processing time and the number of questions that have to be asked on the application form. Reliable customers may be discouraged from applying for credit if they are confronted by a time-consuming form to fill in.

Three approaches to selecting characteristics are commonly used in combination by credit scoring developers:

(1) Stepwise procedures: these represent the most common approach to selecting variables for inclusion in regression models. The initial step involves selecting the characteristic which best explains the variation in creditworthiness using a 1-way analysis of variance. At subsequent steps characteristics can be added to the linear predictors if this leads to a significant improvement in the explained sums of squares. Similarly characteristics can be removed if this does not lead to a significant deterioration. For more details of stepwise regression see, for example, Draper and Smith (1981).

(2) Selecting individual characteristics using the information value:

$$I.V. = \sum_{i,j} (p_{ij} - q_{ij}) w_{ij}$$

where p_{ij}, q_{ij} and w_{ij} are defined above. This is a common measure of the discriminatory power of a characteristic and is used by many scorecard developers in the credit industry. The higher the information value the more the attributes of a characteristic distinguish between the good and bad classes. Typically any characteristic with an information value of over 0.1 will be considered for inclusion in the scorecard.

A similar method of assessing the relationship between creditworthiness and an individual characteristic is to perform a chi-squared test. The calculated chi-squared value is given by:

$$\chi = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_{ij} is the observed number of applicants from class j in attribute i, and E_{ij} is the expected number of applicants from class j in attribute i (assuming uniform proportions of goods across attributes). The sample chi-squared value is then compared with the appropriate point of the chi-squared distribution with $(r - 1)$ degrees of freedom (where r is the number of attributes).

The danger of using one of these two methods is that they only consider the individual relationship of a characteristic with the dependent variable. A characteristic which individually may be uncorrelated with creditworthiness

may be useful in combination with others. For further discussion of measuring the discriminatory power of characteristics, including the information value, see Blackwell (1993).

(3) Expert knowledge of the data: experience of which subsets of characteristics have historically produced the best scorecards can be an effective method of eliminating unpredictive characteristics. This method should always be combined with one of the two approaches described above or there is a danger of perpetuating a sub-optimal set of characteristics. A factor in the characteristic selection also tends to be the need to justify the chosen subset on business grounds.

We choose to use a combination of these methods for selecting characteristics in our analyses. Typically about 16-20 characteristics are selected at the initial vetting stage.

2.3.2 Scorecard construction

A standard approach to constructing scorecards is to fit a linear regression model to the design sample using the chosen subset of characteristics with the weights of evidence as predictor values (we consider a range of other classification techniques that can be used in Chapters 3, 4, 6 and 8). The model takes the form:

$$y = \beta_0 + \beta_1 w_{1j_1} + \ldots \beta_p w_{pj_p}$$

where $y = 0$ if an applicant in the design set is bad and $y = 1$ if an applicant in the design set is good. For a particular applicant, w_{ij_i} represents the weight of evidence for the appropriate attribute j_i of characteristic i. p is the number of characteristics. The parameters β_j are estimated by the method of least squares.

The contribution to the model from attribute k of characteristic j is given by $\beta_j w_{jk}$. The scorecard comes from transforming these "scores" onto an integer scale. For each characteristic the lowest value of $\beta_j w_{jk}$ is transformed to zero. Then the other attribute scores are transformed to integer values, such that the ratio of differences between scores is kept (approximately) constant. The ratio between the maximum and minimum difference in attribute scores is also kept

(approximately) constant across characteristics. The advantage of transforming the scores in this way is that it makes future calculations easier and gives a clearer impression of the relative contribution of each characteristic in the model. We note that the scores are scaled such that the probability of being a good risk has a positive (monotonic) relationship with score.

The scorecard consists of a list of characteristics together with the transformed attribute scores. A future applicant receives an overall score by summing his/her particular attribute scores, which is then compared with an overall threshold score: those scoring above the threshold are accepted for credit, while those scoring below are rejected. For an example of a scorecard see Capon (1982).

2.4 Assessment of scorecards

An important part of the development of credit scoring models is the assessment of performance. We now consider three standard approaches. All three approaches assume that the good and bad applicants in a test set have been allocated a score under some classifier.

(1) The *Divergence statistic*:

$$D = \frac{\mu_g - \mu_b}{\sigma}$$

where μ_g is the sample mean of the distribution of scores for the goods,

μ_b is the sample mean of the distribution of scores for the bads,

σ_g^2 is the sample variance of the distribution of scores for the goods,

σ_b^2 is the sample variance of the distribution of scores for the bads,

N_g is the number of goods in the sample,

N_b is the number of bads in the sample,

and $\sigma^2 = \frac{N_g s_g^2 + N_b s_b^2}{N_g + N_b}$ is the pooled sample variance for the good and bad distributions.

The divergence statistic gives an overall measure of the separability between the distributions of goods and bads. It should be positive because we would expect the goods to score higher than the bads on average (i.e. $\mu_g > \mu_b$). The higher the

value of D, the greater the distance between the sample means and, thus, the more the distributions of goods and bads are separated.

(2) The *Lorenz diagram* and the *Gini coefficient*:

This approach involves comparing the overall cumulative distributions of the goods and bads. Let the cumulative sample proportion good and bad be given by:

$$F(g|s) = N_g(s) / N_g$$
$$F(b|s) = N_b(s) / N_b$$

where $N_g(s)$ is the cumulative number of goods with a score less than or equal to s and $N_b(s)$ is defined equivalently. The Lorenz diagram is a plot of $F(g|s)$ against $F(b|s)$. Figure 2.1 shows an example of a Lorenz diagram for an imaginary classifier.

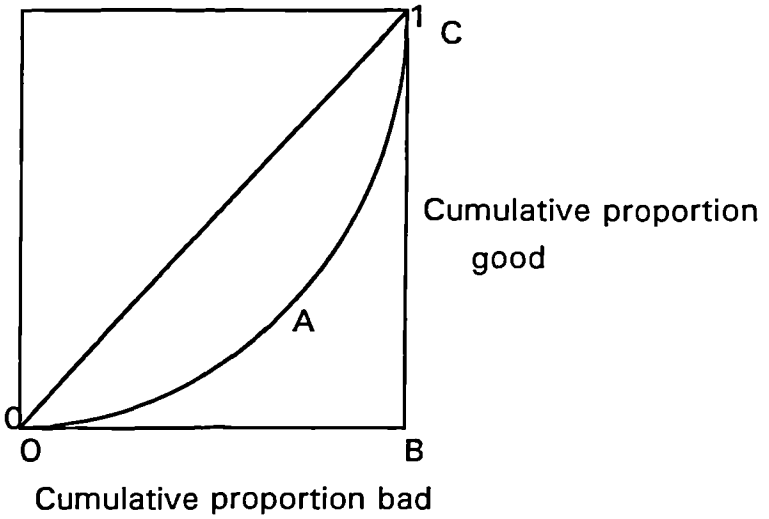


Figure 2.1: A Lorenz diagram.

The Gini coefficient is defined to be:

$$G = \frac{Area(OAC)}{Area(\Delta OBC)}$$

It is used as a standard measure of the discriminatory power of a scorecard by credit scoring practitioners. The higher the Gini coefficient the better the scorecard discriminates between good and bad applicants.

Because the area of the triangle OBC is 1/2, the Gini coefficient can be reduced to:

$$G = 1 - 2Area(OACB).$$

This area can be calculated using the trapezium rule to give:

$$Area(OACB) = \sum_s f(b|s)\{F(g|s-1) + F(g|s)\}/2$$

where $F(g|s-1) = 0$ and $f(b|s) = F(b|s) - F(b|s-1)$. This formula can easily be adapted for non-integer scores.

(3) *Bad rate* and *error rate*:

A different approach to assessing performance of a classifier is to fix a score threshold and to consider the proportion of bad applicants amongst the accepts for the test sample. Alternatively one can consider the error rate, given by the proportion of misclassified good and bad applicants in the full test sample (alternatives to the hold-out method of estimating error rate are considered in Chapter 5). The aim is to produce the scorecard with the lowest bad rate or error rate. In Section 5.3 we develop significance tests to compare the bad rates or error rates from different classifiers. These methods of assessing a classifier allow one to focus on performance in a particular region of the characteristic space, by changing the threshold. The threshold can be chosen to maintain a particular acceptance rate or to fix the expected bad rate.

We have considered three approaches to assessing scorecard performance. The first two methods suffer from one principal weakness: they do not tell you how a scorecard distinguishes between good and bad applicants at individual points, such as the thresholds. (For further discussion of these methods and other similar methods that give an overall measure of the discriminability of a scorecard see Wilkie (1992).)

The third approach does not have this weakness. Another advantage is that it represents objectives that are likely to be of interest to the credit grantor (for example reducing the proportion of bad applicants accepted will lead directly to decreased costs). For these reasons we adopt the third approach in this thesis: our criterion for performance is the minimisation of the bad rate amongst the accepted applicants given a particular acceptance rate. This criterion is unusual for a classification problem and leads to several interesting properties that are discussed in Chapter 5.

Chapter 3

A review of the literature on credit scoring

Eisenbeis (1978) and Crook et al. (1992) describe how the literature on credit scoring can be divided into two groups. First, those papers which deal with the underlying objectives of credit scoring and different aspects of credit granting policy. Secondly, those which consider the relative merits of different classification techniques for constructing credit scoring models (e.g. discriminant analysis, decision trees and logistic regression). The latter papers are considered in Section 3.1 and form a major motivation for the work in later chapters (they can be further subdivided into papers dealing with consumer loans and those dealing with commercial loans). In Section 3.2 we present an overview of the work on credit granting policy and consider other ways in which statistical methodology has been applied in the credit industry (e.g. behavioural scoring, cluster analysis). The emphasis throughout is on identifying areas for further research.

3.1 Review of credit scoring models

As was mentioned in Chapter 1, credit scoring models have largely replaced judgemental methods for assessing credit applicants. This has been a product of the increased demand for consumer credit and the emergence of new computer technology, allowing practical implementation of credit scoring models. The process has been accelerated in the U.S. by the Equal Credit Opportunities Act (1974) and subsequent amendments which have banned discrimination on the basis of sex, race, colour, religion, age and national origin (see Hsia, 1978, for a description of the act). These and other legislative restrictions have encouraged the credit grantor to look for objective methods of identifying good and bad credit risks. Furthermore, general rules were laid down for the proper development of credit scoring systems requiring them to be "demonstrably statistically sound" and "empirically derived" with criteria for sampling and validation.

Chandler and Coffman (1979) present a comparison of judgemental systems with credit scoring models and conclude that, on the whole, the empirical

evaluation process (credit scoring) has no deficiencies not also shared by judgemental evaluation. They also highlight certain advantages of credit scoring not shared by judgemental methods: that empirical methods are based upon actual and not perceived performance and can be statistically validated before implementation; that empirical methods produce more consistent evaluations than judgemental methods; and that empirical methods are more accurate than judgemental evaluation on the average. Having established some of the benefits of adopting a credit scoring system to assess credit applicants, we consider different techniques that have been proposed in the literature.

3.1.1 The use of discriminant analysis and regression techniques for credit scoring

Historically, discriminant analysis and linear regression have been the most widely used techniques for building scorecards. Both techniques have the advantages of being simple conceptually and having widely available routines in statistical software packages. For a theoretical discussion of the two techniques see Sections 4.1 and Chapter 7.

Discriminant analysis in particular has received widespread attention in the credit scoring literature. The first published account of the use of discriminant analysis to produce a scoring system was by Durand (1941). The work was sponsored by the National Bureau of Economic Research (in the U.S.) and produced results showing that discriminant analysis could produce good prediction of credit repayment. (However, it seems that the results were not tested out on new samples to determine the practical value of the scorecards.) Other early studies, such as Myers and Cordner (1957), show that numerical credit evaluation systems can provide good classification performance.

Myers and Forgy (1967) present a comparison of several credit scoring methods: discriminant analysis, stepwise regression, assigning equal weights to each of the predictive characteristics and a series of discriminant analyses applied to subsamples of the original full sample. The subsamples chosen for the last method came from ranking the good and bad risks separately under a scorecard built on the full sample, and then selecting applicants with similar

scores. In particular, this was done for the lowest scoring applicants with the objective of improving discrimination in lower scoring ranges.

The sample used by Myers and Forgy consisted of 600 accepted loan contracts on mobile homes, split randomly into an analysis sample (300 cases) and a hold-out sample (300 cases). Twenty-one characteristics, out of forty-one, were selected for inclusion in the analysis on the basis of a test of significance of good/bad predictiveness. Using the point biserial correlation of actual and predicted scores as an indicator of discriminability gave the surprising result that equal weighting for all 21 characteristics performs best. However, the discriminant analysis on selected cases is shown to be effective at eliminating bad cases in the lower scoring region at the expense of few good cases. The authors conclude that the study indicates there "may be ways in which the basic discriminant analysis approach can be modified to improve its effectiveness, particularly in the case of discrimination at lower score levels, which are of the greatest practical importance in retail credit evaluation."

More recent studies have highlighted some of the methodological and statistical problems that result from using discriminant analysis and linear regression for credit scoring. An examination of the nature of credit data shows that some of the assumptions required for these techniques are violated. (A comparison of linear and logistic regression is presented in Chapter 7). In particular, Eisenbeis (1978) highlights seven general areas of weakness with the traditional approaches to credit scoring (they are based upon points raised in Eisenbeis (1977)). Reichert et al. (1983) consider some of the same issues and attempt to put the theoretical considerations into perspective using empirical results. A sample consisting of 648 applicants for consumer loans from a medium-sized bank was used. Variables which reflect ability and willingness to repay were selected (e.g credit rating, time at present address, relative debt load). We consider each of the points raised by Eisenbeis (1978) in turn with reference to the empirical study described:

(1) *Distribution of the variables*:
The variables used in credit scoring are often of a qualitative nature (or are at least categorical), and so the assumption of multivariate normal data, that is required for discriminant analysis, is violated. However, it can be argued that

discriminant analysis is fairly robust to non-normal data, a view that is backed up by studies such as Myers and Forgy (1963).

One problem is the lack of available tests for multivariate normality; some are discussed by Malkovich and Afifi (1973) although they have not yet been used in the discriminant analysis case. A further problem arises as to how to treat data once non-normality has been established. In Section 4.4 we discuss classification techniques suitable for discrete data. Another approach has been to attempt to transform the data into normal form before estimating the discriminant function (see Pinches and Mingo (1973)). The danger of this type of method is that the transformation may discard irregularities in the structure of the underlying population.

The study by Reichert et al. illustrates some of these points. The data used in the study contained categorical variables that do not satisfy the multivariate normality assumption. This was checked by applying the studentized range test to each of five variables in turn; the test finds the quotient of the data range divided by its standard deviation and looks to see if this falls within a 95% confidence interval. By plotting frequency distributions for the non-normal variables it was found that they were generally skewed to the right for the bads and to the left for the good cases. A common procedure for dealing with non-normal data is to transform it prior to carrying out the analysis (estimating discriminant functions in this case). The appropriate transformations for data skewed to the right are the natural and standard log transformations. A standard log transformation was applied to the data in an attempt to normalize it; as a result 10 out of 15 group variables satisfied the criteria for normality described above.

The next step was to construct three group discriminant models (the three groups were goods, bads and rejects- the justification for including the reject class is examined in Hand and Henley, 1993/4) for the untransformed and transformed data. It was found that the misclassification rate for the transformed data was slightly higher than for the untransformed data. The transformation had little effect on the ability of the model to predict good loans, but it caused a higher percentage of loans to be classified as bad rather than as rejects. This study shows the problems associated with trying to normalize categorical credit data.

(2) *Equality of population covariance matrices* :
When using linear discriminant analysis it is assumed that the covariance matrices for the two populations are equal. If this is not the case then a quadratic discriminant function should be used. Therefore before constructing a scoring model, the developer should test for equality of group covariances. It seems that this has rarely been done for the credit scoring models described in the literature and in almost every case linear discriminant rules were employed.

However, it should be taken into account that quadratic rules are more sensitive to deviations from normality than linear ones. Moreover, Monte Carlo studies by Marks and Dunn (1974) show that linear discriminant rules may give more efficient estimates of the expected error rates than quadratic rules when the group covariances are unequal, if the sample size is small and a relatively large number of variables are used. The results of empirical testing indicate that differences in predictive accuracy between scorecards built using linear and quadratic discriminant functions, can be attributed to two factors: the degree of inequality among the covariance matrices and the difference between group means. Table 3.1 shown below summarises the relationship between these factors and similarity of prediction shown by linear and quadratic rules. S represents similar predictions from linear and quadratic rules whereas D stands for significantly different results.

	Close separation between groups	Wide separation between groups
Low disparity in covariance	S	S
High disparity in covariance	D	S

Table 3.1: comparison of linear and quadratic rules with different types of covariance matrix and separation between groups.

Reichert et al. consider tests of equality of group means and group covariance matrices for their data. It was found that not only were the group means significantly different, but so were the covariance matrices. From Table 3.1

we would expect both linear and quadratic rules to yield similar predictive accuracy; this analysis was carried out and the expected result was obtained, although the quadratic rule showed a greater tendency to classify an applicant as bad.

(3) *The role of independent variables*:
There may be a need for the system developer to demonstrate that each variable contributes significantly to the discriminatory power of the model (to satisfy the business, any court test of the validity of the model etc). This can be achieved for a particular characteristic by using a significance test to compare the performance of the model constructed with all the characteristics included with a model constructed omitting the characteristic of interest.

(4) *Group definitions*:
Most credit scoring systems are constructed using simplistic definitions of credit risk. The most common procedure is to classify a customer as a "bad" risk if they default with repayments for three consecutive months and otherwise to classify them as a "good" risk (as described in Section 2.1). It is assumed that once a customer's true class is determined it cannot change in a subsequent period. Under these assumptions the necessary conditions for using discriminant analysis are satisfied: namely that the populations being considered are discrete and identifiable. However, both these assumptions can be criticised for being unrealistic in practice (as discussed in Section 2.1).

In particular a system where creditworthiness is measured on a continuous scale instead of using a binary good/bad classification might be more appropriate. One such approach is to model the number of payments that an applicant has missed, although this gives an integer rather than a continuous measurement. This is an area requiring further research. More ambitiously a credit grantor might be interested in measuring creditworthiness on a multivariate scale.

(5) *Use of inappropriate a priori probabilities and costs of misclassification*
When performing linear or quadratic discriminant analysis one has to specify prior probabilities of an applicant belonging to the good and bad classses. It is also necessary to specify the relative costs of misclassification for the two classes. It appears from the literature that little emphasis is given to choosing the a priori probabilities (often they are assumed to be equal) and that the costs

of misclassification for goods and bads are assumed to be equal. This means that the misclassification rate may be a poor estimate of the true error rate when the discriminant rule is applied to the applicant population. If the sample used in the analysis is a random sample drawn from the full population of applicants, then using the sample proportions as estimates of the population priors will be appropriate.

Reichert et al. (1983) estimate population priors from a large sample from the full population of applicants (which includes the analysis sample) and find their new estimates to be significantly different from the analysis sample priors. They then proceed to construct discriminant models using a) equal priors, b) sample priors and c) the estimated population priors. The overall predictive accuracy of methods b) and c) was found to be the same and superior to that of a). However, method c) involved an increase in type II errors and a decrease in type I errors over method b). This means that the estimated population priors lead to a model that achieves higher accuracy than the sample priors model in predicting goods, but lower accuracy for the bads and rejects. It does seem more appropriate to use the population priors, but it is not clear without examining the relative costs of misclassification whether it is beneficial in the example mentioned above.

(6) *Estimation of classification error rates*:
Eisenbeis (1978) makes the assumption that the error rate is the most appropriate measure of performance (but the discussion applies equally to our criterion of minimising the bad rate amongst accepted applicants). The different approaches to estimating the error rate are reviewed in Hand (1986) and Toussaint (1974).

If the original analysis sample is used to estimate the expected error rate (the reclassification approach) then biased results are obtained. However, the majority of developers of credit scoring systems appear to use the hold-out method. This involves randomly dividing the available sample into an analysis sample, which is used to perform the discriminant analysis, and a hold-out or test sample which is used to validate it.

Reichert et al. (1983) show that, given their data set, using the reclassification approach to estimating error rates leads to a bias of about 2-3 % compared to

using a 50% random holdout procedure. They then demonstrate that using the *holdout procedure* leads to reasonably consistent and reliable discriminant coefficients, by showing that the actual coefficients and relative coefficients (obtained by dividing the coefficients for each variable by the coefficient for the most predictive variable, credit rating) have almost the same degree of stability based on 24 random 50% holdout samples.

Further analysis showed that the holdout procedure generates reasonably consistent and reliable coefficients for characteristics with significant discriminatory power. It was also found that the holdout procedure gives stable mean error rates over a range of holdout percentages. An advantage of the holdout method is that it is computationally much simpler than alternative methods like the Lachenbruch procedure (and they give very similar results for large samples). The authors concluded that, unless the sample size is extremely small, the standard holdout method can prove to be a suitable testing procedure. (In fact, the dimensionality of the characteristic space can also influence the reliability of the holdout procedure.) The size of our samples are large (see Table 2.1) and, thus, we choose to adopt the holdout method in this thesis. A more detailed discussion of assessing scorecard performance and a comparison of methods of error rate estimation is presented in Section 2.4 and Chapter 5.

(7) *Selection of analysis samples*
One additional problem that is associated with building credit scoring models concerns the selection of representative analysis samples. The sample used to construct the scoring system often only consists of accepted applicants. This is because the true creditworthiness of the rejected applicants is usually unknown (although, for the work described in this thesis, we do in fact have access to a sample with the true creditworthiness known for the rejects). The result can be bias in the scorecard. This provides a fundamental motivation for performing reject inference. This issue is explored in depth in Chapter 6.

3.1.2 Other credit scoring models

Because of the statistical problems associated with the use of discriminant analysis to build scorecards outlined in the previous section, there has been a considerable amount of research devoted to identifying more suitable

classification techniques for credit scoring data. This began with consideration of standard statistical methods such as logistic regression and has moved on to include the application of techniques from other fields, such as expert systems, neural networks and genetic algorithms from computer science. We consider a range of empirical studies covering most of the classification methodologies that have been applied to the credit scoring problem:

(1) Chatterjee and Barcun (1970):
This is an early example of the application of a nonparametric approach to credit scoring. The authors propose classifying an applicant to the class with which it has most in common, this being done so as to minimise the expected loss from misclassification. It is a variant of the "closest neighour" rule given by Hills (1967) (and a simplified version of the nearest neighbour rule described in Section 8.2.1). A jackknife method is used to estimate the classification error rates. Results are presented for data consisting of personal loan applications made to a bank in New York. The effect of varying the relative costs of misclassification is investigated. Unfortunately, no attempt is made to compare the nonparametric classification rule with other credit scoring methods, such as discriminant analysis. This would have enabled a more objective assessment of the proposed methodology.

(2) Wiginton (1981):
This represents one of the first published accounts of the application of a logit model (logistic regression) to the credit scoring problem. Logit models are more appropriate for credit scoring than techniques like discriminant analysis, because they do not assume that the characteristics are drawn from a multivariate normal distribution. They also allow the assumption that the dependent variable (creditworthiness) has a binomial distribution, the appropriate distribution for binary data. (We consider the theoretical differences between linear and logistic regression in more detail in Chapter 7).

An empirical study to compare the performance of the logit model with discriminant analysis is presented using data collected by a major oil company. Only three characteristics were used in the scoring models (in indicator form): "years at present employment", "living status" and "occupation type". The results showed that discriminant analysis gave identical performance to a classifier based upon chance (by allocating all cases to the largest group). In

comparison the logit model achieved a 62% correct classification result compared with an expected 58% correct classification using chance.

The implication of the results is that, although the logit model outperforms discriminant analysis, neither model is very useful for making classification decisions with the given dataset. However, a couple of questions can be raised about aspects of this study: first, the number of characteristics used was unrealistically small. It might be possible to achieve better discrimination with a larger number of variables (this appears to have been a limitation of the dataset rather than the method of variable selection). Secondly, by changing the threshold for the discriminant analysis it should, in theory, be possible to produce a lower misclassification rate than the chance classifer.

Wiginton concludes by saying that credit grantors, such as the oil company who supplied the data in this study, should shift the burden of assessing credit applicants to specialized institutions, such as banks, who have access to more data and expert technical knowledge. He argues that the cost of this service may be less than the total cost of running a credit and collections department for consumer accounts within the firm. This seems to be at odds with current trends in the credit industry.

(3) Grablowsky and Talley (1981):
The authors make a comparison of probit analysis and linear discriminant analysis for credit scoring. A probit model is an example of a generalised linear model (see McCullagh and Nelder (1983)), first developed by Finney (1952) for the analysis of toxicology problems. It assumes that there is an underlying threshold, $\bar{l}_j$, on a linear combination of the predictor variables, l_j, for the jth individual, such that if $l_j > \bar{l}_j$ the jth individual is classified as a good and if $l_j < \bar{l}_j$ the jth individual is classified as a bad. The probit model was thought to be more suitable for credit scoring than discriminant analysis because, rather than assuming a multivariate normal distribution for the response, it assumes that the underlying thresholds $\bar{l}_j$ have a normal distribution. It also constrains the predicted response $\hat{P}(g|\mathbf{x})$ to lie between 0 and 1. For more details of the probit model see Section 4.7.

The dataset used in this study comes from a large, midwestern retail chain in the U.S. Eleven explanatory characteristics were used that had been found to

be predictive over a ten-year period. Using a stepwise discriminant procedure four of the characteristics were included in the model. For the probit analysis the coefficients are unique and a likelihood ratio test can used be to test their individual significance. Nine characteristics were found to be significant using this procedure and thus included in the model. The resulting classification table is shown in Table 3.2 in order to clarify a comment on the results made below.

True class		Proportion classified as Good risk	Bad risk	Total
Good	Discriminant	1.00	0.00	1.00
	Probit	0.91	0.09	1.00
Bad	Discriminant	0.20	0.80	1.00
	Probit	0.07	0.93	1.00

Table 3.2: classification results from the comparison of discriminant analysis and probit analysis from Grablowsky and Talley (1981).

The authors state that, if the misclassification of the bad risks is more costly than the misclassification of the good risks (excluding computation costs), then the probit analysis has achieved superior performance. However, if we were to alter the acceptance threshold for the discriminant analysis such that the correct classification of the goods is lowered to the level achieved by the probit analysis (91%), then we would see a corresponding improvement in performance for the bads. This might even be sufficient to increase the proportion of bads correctly classified to the level 93% achieved by the probit analysis. Therefore, on the basis of these results, it is not possible to conclude which technique is achieving the best performance. In our comparison of classification techniques we choose to fix an acceptance threshold for all techniques to provide a fairer comparison. Nevertheless these results do indicate that a probit classifier is a viable alternative to discriminant analysis, assuming that the computation costs are not prohibitive. In fact, the probit model usually gives classification results that are very similar to those from logistic regression and the latter is more popular nowadays.

(4) Srinivasan and Kim (1987):

The authors carry out a general comparison study of several credit scoring systems. Four statistical models were used: linear discriminant analysis,

logistic regression (logit), goal programming (GP) and the recursive partitioning algorithm (RPA - often referred to as decision trees). A judgemental model based upon the Analytic Hierarchy Process (AHP) was also used. All of the statistical methods are described in Chapter 4. They represent a range of standard parametric and non-parametric classification methods that have been applied to credit scoring. The authors anticipated that the RPA would lead to better discrimination than simultaneous partitioning procedures like MDA, logit and GP. This is because simultaneous partitioning procedures make the implicit and unrealistic assumption of convexity of the class regions in the characteristic space.

The data used in this study consists of a sample of accepted commercial applicants for credit. The objective was to to replicate a loan officer's risk rating of loans, after they had been accepted. Because the study was concerned with commercial loans, the data consisted of less variables (8) and smaller samples (215 cases) than are typically available for consumer loans. Bootstrapping (see Section 5.2.1.2), a technique that is especially appropriate when the data violate distributional assumptions of the classification method, and the holdout method were used to estimate the error rates. The results showed that the RPA did indeed give slightly superior performance. Of the other techniques, the Logit model gave marginally better performance than the discriminant analysis and GP.

In a discussion of the paper, Eisenbeis makes some criticisms of the study. He argues that one can't tell whether the small differences in performance between techniques are due to differences in the robustness of techniques when the underlying assumptions are violated or whether they are truly due to superior or inferior performance. He suggests that, because of the small sample sizes, a Monte Carlo experiment drawing samples from known distributions and populations with known parameters would be more appropriate. However, in defence of the original paper, it should be pointed out that comparison studies using real datasets are of more practical interest than studies using hypothetical populations.

(5) Boyle et al. (1992):

This paper presents a comparison of credit scoring techniques, similar to that conducted by Srinivasan and Kim (1987), using consumer credit data. The

statistical methods discussed are discriminant analysis, the recursive partitioning algorithm (RPA) and hybrid methods which use both techniques. To avoid the problem of sample selection bias, the scoring system was designed to identify slow payers amongst existing credit card holders. The dataset consisted of 1001 accounts with 24 standard application characteristics.

The best performance on a holdout sample was achieved by a hybrid classifier. This classifier used RPA to construct decision trees to identify interactions between the original characteristics. The hybrid method involves constructing two decision trees in this way and including them in a linear discriminant model as characteristics. The decision tree characteristics were found to be by far the most important characteristics in the discriminant function.

It was also found that the standard linear discriminant method performed slightly better than the standard RPA (for the test sample). In conclusion the authors commented that discriminant analysis and decision trees have complementary strengths that can be incorporated to give a successful hybrid method. In particular, the strength of discriminant analysis is that it uses all of the data in building the model and the strength of the RPA is that it enables the modelling of complex dependencies between variables.

(6) Leonard (1988) and (1993):
Leonard (1993) summarises the findings of his PhD dissertation (1988) on the application of random effect logit models to commercial loan applications. The objective of the study is to simulate the decision process of commercial loan officers. The random effects model is used to enable the estimation of the effect of the branch making the loan. This innovation is needed because there are multiple branches and data is only available on a small sample of the branches. By including the branch parameters as random variables, this allows inferences from the model to be extended to the entire population of branches. Another problem that arises is for branches where there are only a small number of loan applications made, the parameter estimates for the branch effect using a fixed effects model can be extremely high (in some cases even infinite). By assuming the branch effect to be random, the parameter estimates for the branch effect shrink towards an overall central estimate for the branches where the sample information is scarce.

A comparison study of the logit model with fixed and random effects and a linear discriminant model is presented. The method of error estimation was to take repeated random holdout samples from the original sample. This procedure was repeated 5 times and then the results were averaged. (We adopt a similar approach to assessing performance in Chapter 8). The classification results showed that the random effects logit model outperformed the corresponding fixed effects model and discriminant analysis. This study indicates the potential of the random effects model for use in credit scoring problems.

(7) Fogarty and Ireson (1993/4):
In this paper a genetic algorithm is used to optimise a Bayesian classifier using a set of examples. Genetic algorithms are sophisticated computing methods based upon a biological metaphor of evolution (see Holland (1975), Goldberg (1989), South et al. (1993)). They were thought to have potential for competing with standard credit scoring methods because they are known to efficiently search large solution spaces where conventional statistical theory is inappropriate. In Section 4.11 we briefly discuss the principles behind the genetic algorithm and consider its properties.

The authors describe the IDIOMS system, a software environment for implementation of the genetic algorithm. They present a comparison of the genetic algorithm with the default rule, a decision tree (a new version of ID3 using entropy), a Nearest Neighbour algorithm (see Chapter 8) and a simple Bayesian classifier (independence model). A large data set consisting of 51,020 accepted applicants for consumer credit (94% good and 6% bad) was randomly split into ten equal test sets. At each stage a test set was selected and classifiers constructed on the remaining dataset. The classifiers were then used to estimate performance measures on the test set and the results averaged over test sets. Three measures of performance were used: classification accuracy (the proportion of goods and bads correctly classified), a simple measure of profitability and acceptance rate.

The mean classification accuracy, profitability and acceptance rates are shown for each classifier in Table 3.3. Using the default rule (classifying all applicants as good) gave very high overall classification accuracy of 94%. In fact, none of the other classifiers were able to achieve the same level of

accuracy. However, the genetic algorithm (92.74%) did perform better than the other three, with the simple Bayes rule (89.05%) performing the worst.

	Accuracy	Profit	Acceptance
Default	94.022	49.082	100.00
New ID	91.202	50.783	90.386
NN	91.009	57.647	74.010
Bayes	89.046	66.153	74.400
IDIOMS	92.743	68.804	79.774

Table 3.3: Measures of performance in the comparison study by Fogarty and Ireson (1993/4).

The profitability criterion used was given by the number of goods correctly classified minus eight times the number of bads incorrectly classified all divided by the total number of goods. Because this gives a weighting of 8 to the bad applicants it removes the advantage of the default rule. The results for profitability show that the genetic algorithm gives the best performance, followed by the simple Bayes classifier. The NN rule outperforms the decision tree and the default rule.

An explanation for the differences in classifier ranking between the classification accuracy and profitability results is provided by considering the acceptance rates. Any classifier having a relatively low acceptance rate is likely to have a relatively higher profitability rank than classification accuracy rank, because of the importance the profitability criterion gives to bad applicants (and the proportion of bad applicants accepted will decrease rapidly as acceptance rate decreases). For this reason, although the simple Bayes classifier gives poor classification accuracy, it gives the second best profitability. Using a similar argument, the decision tree gives similar performance to the NN rule for classification accuarcy, but appears significantly less profitable because it has a higher acceptance rate. Despite these comments, it does appear from the results that the genetic algorithm peforms well giving the best profitability and the closest accuracy to the default rule.

3.1.3 Summary and implications for future research

In this section we have reviewed some of the important contributions to the literature on the appropriate classification techniques to use for credit scoring models. We split the review into two parts: first, a discussion of the traditional approach, discriminant analysis, with particular emphasis on methodological and conceptual weaknesses; and, secondly, a discussion of papers proposing new classification methodologies.

Eisenbeis (1978) and Leonard (1988) describe a number of papers that have not yet been mentioned that deal with traditional credit scoring methodology. Work on consumer loans includes Smith (1964), Orgler (1971), Apilado et al. (1974), Chandler and Coffman (1983/4) and Overstreet and Kemp (1986). Work on commercial loans includes Altman (1968), Orgler (1970), Edmister (1972), Blum (1974) and Doreen and Farhoomand (1983). Leonard (1988) uses a tree diagram to illustrate different aspects of the literature on commercial loans.

The work on alternative classification methodologies described in Section 3.1.2 indicates that there is potential to develop scoring models that can outperform discriminant analysis. Logistic regression is becoming increasingly popular with scorecard developers and a comparison with linear regression is considered in Chapter 7. Other papers which include the application of logistic regression to credit scoring are Steenackers and Goovaerts (1989) and Gilbert et al. (1990).

The recursive partitioning algorithm (decision trees) has been used extensively and gave good performance in studies by Srinivasan and Kim (1987) and Boyle et al. (1992). Other work on the application of decision trees and machine learning algorithms to credit scoring is described by Carter and Catlett (1987) and Davis et al (1992). In Section 4.9 we describe decision trees in more detail and consider two extensions: fanned trees and decision graphs.

Fogarty and Ireson (1993/4) have shown the potential of the genetic algorithm for building credit scoring models. Other techniques such as neural networks (see Section 4.10) have received less rigorous testing but, nevertheless, have the potential to provide models that can outperform discriminant analysis because of their ability to fit non-linear relationships.

In this thesis one of our objectives is to search for new and appropriate approaches to building credit scoring models. This motivates a theoretical comparison of classification techniques in Chapter 4 and the application of the k-Nearest Neighbour method in Chapter 8. To put this objective into perspective, we note that if the data has poor separability then it may not be possible to achieve significant improvements in performance using any classification technique. Moreover, the search for more predictive characteristics may be the only way to make a large difference to the performance of a credit scoring model.

3.2 Aspects of credit granting policy

In the last section we considered different approaches to estimating the probability of an applicant being a good risk given the characteristic vector, denoted by $\hat{P}(g|\mathbf{x})$. The models described are usually constructed with the aim of minimising the bad debt rate (or the classification error rate). Eisenbeis (1978) argues that much of the literature is incomplete because it does not take into account other dimensions of the credit granting decision, in particular profit maximisation. We now consider the consumer credit granting problem from a more general perspective and examine ways in which estimates of applicant risk can be incorporated into an optimal credit granting policy. Other ways in which credit scoring methodology and statistical theory can be applied in the credit industry are discussed in Section 3.2.2.

3.2.1 Strategies for credit granting

An early study of the problem of formulating an optimal credit granting strategy was Greer (1967). He focused on selecting the optimal number of contracts to be accepted by maximising "credit-related profits", which consist of the sum of the present values of the (1) profit from credit sales in the current period, (2) profit from credit sales in future periods and (3) profits from cash sales in both the current and future period. Greer's model does not on its own provide an accept/reject rule for an individual applicant. However, by fixing the acceptance rate using Greer's model, one is able to get a bound on the

maximum probability of default that should be accepted for an individual applicant. In this way it can be combined with a standard scoring model to take decisions on individual applicants. Another positive aspect of the model is that it is a multi-period model, because it includes the future value and additional businesses that might be gained as a result of granting a particular number of loans.

Mehta (1968,1970) addresses the issue of how to make a decision on an individual applicant. A sequential decision process is described that rests on two premises: (1) information is costly and not all information is needed to take a decision on a particular applicant and (2) past experience can be a reliable way to predict the future credit performance of an individual. Three types of cost are considered in the model:

(1) Acceptance cost = (Probability of non-payment).(variable product cost) + average investment cost + average collection cost.
(2) Rejection cost = (Probability of payment).(contribution margin)
(3) Cost of further information = cost of acquiring additional information + cost of additional decision.

For each given amount of information available, the credit grantor can choose to accept, reject or collect more information on a particular applicant. The strategy is chosen which minimises the expected cost.

In his second paper, Mehta extends the model to include optimal decisions for collection of outstanding debts. This includes the value of any resulting future sales or credit requests. A Markov process is used to estimate the costs of collecting accounts i periods old.

Bierman and Hausman (1970) propose dynamic programming models which incorporate and allow revision (based upon repayment performance) of subjective prior estimates of the probability of default for a new customer. It is assumed that loans are made for one period only. At the end of each period a new loan can be requested. The outcome from each period is a Bernoulli trial (with probability p of collection) and the cumulative outcomes over n time periods is a binomial process. The prior distribution for p is assumed to be a Beta distribution with parameters n and r, where r is the number of collections. The parameters n and r increase as more loans are made and repayments

received. At each stage the expected monetary return is compared with the expected monetary cost leading to an accept/reject decision. Further work on these models has been carried out by Dirickx and Wakeman (1976) and Srinivasan and Kim (1987). However, their practical usefulness is limited because of the restrictive assumption that credit is extended for one period only.

An interesting multi-period approach to modelling credit behaviour is taken by Cyert et al. (1962). They model repayment behaviour using stationary Markov chains. It is assumed that in each month an account can belong to one of (n+2) states: paid, current, one month overdue, ..., (n-1) months overdue, and bad debt. The probability that an account in one state in one period will move into a particular state in the next period is defined by a transition matrix of probabilities. The elements of this matrix are estimated using maximum likelihood estimation. Cyert and Thompson (1968) propose using separate transition matrices for different risk classes. Credit limits are then calculated to maximise a profit criterion over n periods. Further work on the application of stationary Markov chains is described by Corcoran (1978) and van Kuelen, Spronk and Corcoran (1981).

Frydman, Kallberg and Kao (1984) make a comparison of stationary and nonstationary Markov chains and an extension, the mover-stayer model. The mover-stayer model assumes that the population consists of "stayers", individuals who stay in their initial state, and "movers", individuals whose repayment behaviour follows a stationary Markov chain. Residual matrices were used to highlight a weakness of the Markov chain models. It was found that the predicted transition matrices seriously underestimated the diagonal entries of the corresponding observed matrices. They conclude that the mover-stayer model provides a much better description of the data than either stationary or non-stationary Markov chains.

This methodology could be extended to take into account the characteristic vectors for credit applicants, as well as the different classes. As a result one could use this type of stochastic system to make an accept/reject decision on an individual applicant. It has the advantage over conventional credit scoring techniques that it gives a probability of belonging to a range of states after a fixed period of time, rather than a simple good/bad prediction. It may be of considerable benefit to the credit grantor to be able to distinguish between the

length of time that an applicant will default for. In Section 4.11 we consider Markov models in more detail and examine how the transition matrix can depend upon the characteristic vector.

Edmister and Schlarbaum (1974) consider the problem of selecting the optimal credit granting system. They propose choosing the method of analysis which maximises the difference between the expected net value of granting loans, given a fixed number of applications and method of analysis, and the expected net value without using the analysis. Eisenbeis (1978) points out that their procedure is non-optimal because they do not include endogenous components, such as interest rates and collections policies, in their model. Long (1976) extends the Edmister and Schlarbaum model to include (1) startup and updating costs, (2) changes in system efficiency over time and (3) changes in the number of applicants.

Oliver R.M. (1992) proposes a model for profitable selection of individual scored applicants for credit. The model works by splitting the risk score line into accept/reject regions depending upon the expected profits at each score. An investigation into the effect of score-splitting policy upon expected profit is carried out using three possible score regions, S_i for $i = 1,2,3$. Two policies are considered: first, an applicant with score S is accepted if $S \geq S_1$ and is rejected if $S < S_1$ (this represents the standard approach); and, secondly, an applicant with score S can be accepted or rejected for $S \geq S_1$, depending upon the expected profit in the two regions, and must be rejected if $S < S_1$. The results showed that, in this simplified case, the score-splitting policy can lead to increased expected profits depending upon the risk probabilities. Surprisingly, it was found that the optimal score-splitting policy could lead to rejection of a higher-score applicant and acceptance of a lower scoring applicant.

Having considered various aspects of credit granting policy we mention four key elements identified by Eisenbeis (1978).

(1) The credit granting process is a multi-period problem. The decision to grant credit in one period can also affect the value of the customer relationship over future time periods. Many of the models considered in this section, such as the Markov chain models and the approach of Edmister and Schlarbaum (1974), take these factors into account.

(2) Some authors, such as Mehta (1968,70) and Bierman and Hausman (1970), emphasise the need to take into account the cost of information in constructing a scoring system.

(3) Most of the credit policy models considered do assume a system of estimating probabilities of applicants defaulting (or opportunity costs of granting loans). In other words, they incorporate the types of credit scoring models that we focus on in this thesis.

(4) Some of the methods considered, such as those of Bierman and Hausman (1970), allow the probabilities of default to be readjusted over time to take into account customer performance. An alternative approach to assessing customer performance, in order to make new credit granting decisions, is to use behaviour scoring (see Section 3.2.2).

3.2.2 Other statistical applications to the credit granting process

(1) Behavioural scoring:
Behaviour scoring is the process of determining the probability that a credit account will remain in, or return to, a "good" state. It is clearly of paramount importance that a credit manager is able to predict the performance of existing accounts, whether to anticipate problems with repayments or to successfully market alternative financial products. As a result behaviour scoring can be a very important part of an overall credit granting strategy.

The standard credit scoring techniques used for building application scorecards can be used to build behaviour scorecards (as described in Chapter 2). However, there are some aspects of the data and the objectives of behaviour scoring that call for different considerations: first, the characteristics available for behaviour scoring are often highly correlated, because the performance data is all based upon one account; and, secondly, the purpose of the behaviour scorecard may be to perform some function, such as setting a credit limit, which does not require a simple accept/reject classification.

Blackwell and Sykes (1992) address the second point mentioned above. Using a traditional scorecard approach, the optimal credit granting strategy would appear to be to grant unlimited credit to applicants scoring above the threshold and to grant nothing to those scoring below. However, it is clearly inappropriate for a financial institution to adopt such a policy. Furthermore, this approach does not enable differentiation between customers who pay back a large proportion of their balance and those who pay back nothing.

The concept of marginal risk is introduced in order to allow the formulation of an optimal credit limit strategy. The marginal risk ($M(x)$) is defined to be the expected proportion of the last £x of balance that is paid off in a given time period. The authors demonstrate that for a given behaviour score, although the overall risk is fairly constant, the marginal risk varies according to the outstanding balance. A model which summarises the relationship between marginal risk, average balance and behaviour score is proposed. This model gives contours of equal marginal risk for different behaviour scores and average balances. The optimal credit limits come from the contour which gives expected profits of 0. A sensitivity analysis of the parameter x showed that the optimal credit limits are quite sensitive to choice of this parameter. Despite some practical problems with the approach taken in this paper, it does provide a useful method of selecting credit limits. For further discussion of behavioural scoring see Chandler and Coffman (1983/4).

(2) Cluster Analysis:

Another example of the application of standard statistical methodology to the credit granting problem is the paper by Lundy (1992). Cluster analysis is used to identify distinct subgroups of the population. The main application of the results is likely to be the identification of suitable sub-groups of the population for marketing purposes e.g. selecting customers to include in a particular mailing. The clustering technique used is an optimisation method in which:

(1) The number of clusters is fixed at the start.

(2) Applicants in the sample are reallocated so as to minimise the within cluster sum of squares.

(3) The procedure terminates when no further significant reduction in the sums of squares can be achieved.

Six clusters are identified which correspond well to distinct sub-groups in the population (such as "School leavers", "Professionals", "Pensioners" etc). It was also observed that the different sub-groups had different proportions of good and bad applicants. Although this was not an objective of the cluster analysis, it appears to have achieved reasonable discrimination between the good and bad classes. This study illustrates that statistical methodology can be used to tackle other aspects of the credit granting process. A second application of cluster analysis to the credit granting procedure is described by Edelman (1992).

Chapter 4

Statistical classification techniques for credit scoring

In this chapter we review statistical methods for classifying cases to one of a number of distinct populations. A range of parametric and non-parametric techniques are considered, some of which are currently used by developers of credit scoring models. In particular we focus on selecting methods to estimate the probabilities of applicants for credit belonging to the good or bad classes, based upon their application characteristics ($\hat{P}(g|\mathbf{x})$ and $\hat{P}(b|\mathbf{x})$). Our aim is to identify classification methods which are appropriate for credit data and can provide good discrimination between classes.

We distinguish between direct and indirect methods of estimating $P(g|\mathbf{x})$. In the first approach, of which logistic regression is an example, we estimate $P(g|\mathbf{x})$ directly as a function of some parameters. In the second approach, of which discriminant analysis is an example, we estimate $P(\mathbf{x}|g)$ and $P(\mathbf{x}|b)$ separately and then derive $P(g|\mathbf{x})$ using Bayes' theorem:

$$P(g|\mathbf{x}) = \frac{p(\mathbf{x}|g)P(g)}{p(\mathbf{x}|g)P(g) + p(\mathbf{x}|b)P(b)} \qquad (4.1)$$

where $P(g)$ and $P(b)$ are the population priors for the good and bad populations, respectively.

Dawid (1976) refers to the indirect and direct methods of estimating $P(g|\mathbf{x})$ as the sampling paradigm and diagnostic paradigm approaches, respectively. He points out that a sampling fraction which varies across the characteristic space will lead to distortion of methods which use the sampling paradigm, but not of methods which use the diagnostic paradigm. Thus, if a sample selection mechanism is being used such that an applicant's characteristic vector affects their chances of being included in the sample, it is more appropriate to use a direct method of estimating $P(g|\mathbf{x})$. This has important implications for selecting a method of reject inference (see Chapter 6 and Henley and Hand (1993/4)).

A common approach to formulating a classification rule (see Hand (1992)), based upon the estimated class probabilities $P(g|\mathbf{x})$, is to classify an applicant

to the class for which the expected cost of misclassification is minimised (the Bayes minimum risk classification rule). However, as has already been mentioned in an earlier chapter, we adopt an unusual performance criterion in this thesis, namely the minimisation of bad rate amongst the accepts (see Chapter 5). The relative costs of misclassification could be included in our performance criterion if suitable values were known.

4.1 Discriminant analysis

4.1.1 Classical linear discriminant analysis

This has traditionally been the most popular method for building credit scoring models described in the literature. It is one of the earliest statistical techniques for discriminating between distinct populations and was first introduced by Fisher (1936).

The method uses an indirect approach to estimating $P(g|\mathbf{x})$. Adopting the notation of Hand (1992), the method assumes that the conditional distributions for the goods and bads, $P(\mathbf{x}|g)$ and $P(\mathbf{x}|b)$, are ellipsoidal (a class which includes the multivariate normal) and have common covariance matrix Σ and mean vectors μ_i (i = 1 (good), 2 (bad)). The optimum linear discriminant function is defined to be the function $\mathbf{a}^t\mathbf{x}$ (where $\mathbf{x}$ is the characteristic vector), such that $\mathbf{a}$ maximises the standardized distance between the group means given by:

$$\frac{\mathbf{a}^t(\mu_1 - \mu_2)}{\sqrt{\mathbf{a}^t \Sigma \mathbf{a}}}.$$

The resulting optimum formula for $\mathbf{a}$, estimated from the data using maximum likelihood estimation (MLE), is:

$$\hat{\mathbf{a}} = \hat{\Sigma}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2).$$

Classification of individual applicants involves imposing a threshold on the linear discriminant function.

The linear discriminant approach to credit scoring is appealing because of its conceptual simplicity and its wide availability in statistical software packages. However, as we saw in Section 3.1, credit scoring data do not generally satisfy

some of the basic assumptions required. In particular, the assumption of ellipsoidal distributions for the goods and bads may be unrealistic. Because of the discrete nature of most characteristics, a multinomial distribution would be more appropriate. Reichert et al. (1983) also demonstrated that the assumption of equal covariance matrices is unlikely to hold in practice.

4.1.2 Quadratic discriminant analysis

Quadratic discriminant analysis addresses one of the potential weaknesses of the linear discriminant method by relaxing the condition of equal covariance metrices. This allows the separating surface between the class distributions to take a more complex quadratic shape.

Despite the greater flexibility of the quadratic discriminant rule, comparisons of linear and quadratic discriminant rules using credit scoring data have shown that there is very little difference in performance between the two techniques (e.g. Reichert et al. (1983)). More general studies of the linear discriminant rule have also shown its robustness relative to the quadratic rule (e.g. Aitchison et al. (1977) and Titterington et al. (1981)). Reasons for this are that the quadratic rule involves estimating many more parameters and is vulnerable to overfitting of the data. We conclude that the quadratic approach does not appear to offer any advantage over the classical linear discriminant approach.

4.1.3 Regularized discriminant analysis

A potential problem with using linear discriminant analysis is highlighted by Friedman (1989). This is that the discriminant function is heavily dependent upon the smallest eigenvalues of the covariance matrix, Σ. As a result, a large part of the variation in discriminant scores arises from characteristics with small sample variances. Friedman proposes using a regularized form of discriminant analysis, whereby the covariance matrix is shrunk away from the sample estimate to some other value using a shrinkage parameter. It is pointed out that linear discriminant analysis can be considered as a regularized version of quadratic discriminant analysis, where the class covariance matrices are shrunk to their average value. Because the methodological weaknesses of using linear

discriminant analysis for credit scoring apply equally to the regularized version, this does not seem an appropriate technique to adopt.

4.2 Linear programming

Linear programming involves the minimisation or maximisation of a linear function subject to specified linear constraints. Freed and Glover (1981) propose the application of linear programming techniques to discrimination problems. Although the fitted model takes the same form as the linear discriminant function, we discuss linear programming and linear discriminant analysis separately because they are treated separately in the credit scoring literature.

Srinivasan and Kim (1987) outline linear rules for the two group discriminant problem, in the general case where it is not possible to completely separate the two classes. If applicants have characteristic vectors $\mathbf{x}_i$, where $i = 1...n_1$ corresponds to good applicants and $i = n_1 + 1...n$ to bad applicants, then a suitable linear rule involves estimating the parameter vector a and a slack variable s to satisfy the following conditions:

$$\text{Min} \sum_{i=1}^{n} s_i$$

such that

$$a\mathbf{x}_i \leq b + s_i \qquad i = 1...n_1$$
$$a\mathbf{x}_i > b - s_i \qquad i = n_1 + 1...n$$

where b is any positive constant. The slack variable, s, is needed to take account of non-separability and permit feasible solutions. The above formulation constrains s according to the individual observations. Srinivasan and Kim also consider the more restrictive case of constraining s according to class (having one s component for each class).

Classification results presented by the above authors showed that the linear programming rules gave comparable performance to linear discriminant analysis, but did less well than logistic regression and the RPA. We conclude that linear programming methods can provide robust credit scoring models. However, by their nature, they are unable to model complex non-linear

relationships (or interactions between characteristics) and they do not appear to offer any significant advantages over linear discriminant analysis.

4.3 Independence models

A simple (and perhaps simplistic) approach to estimating $P(g|\mathbf{x})$ is to assume that the characteristic values are independent given the applicant class. The resulting Bayes independence model for p characteristics is given by:

$$p(\mathbf{x}|g) = p(x_1|g)p(x_2|g)\ldots p(x_p|g)$$

where x_i is the ith component of $\mathbf{x}$. A similar relation holds for $P(\mathbf{x}|b)$ and Bayes theorem is used to combine them (as in equation 4.1) to estimate $P(g|\mathbf{x})$.

In most credit scoring applications the independence assumption is unlikely to hold. However, despite the inappropriateness of this assumption for most medical applications, Hand (1992) reports that the independence model is widely used in the field of medical diagnosis and often does surprisingly well. In a comparison of discrimination techniques for ordered categorical data by Titterington et al. (1981), a smoothed independence model is used to estimate the class conditional densities given by:

$$p(\mathbf{x}|i) \propto \left\{ \prod_{r=1}^{p} \frac{n_i(x_r) + 1/C_r}{N_i(r) + 1} \right\}^B$$

where $n_i(x_r)$ is the number of applicants in class i with attribute x_r, C_r is the number of attributes of characteristic r, $N_i(r)$ is the number of applicants in class i with characteristic r not missing and B is an association factor representing the "proportion of non-redundant information" in the variables. The results show that this independence model is robust and gives good classification accuracy.

Hilden (1984) discusses why the independence model can often perform well when the independence assumption is not met. One of the explanations put forward is that the independence model requires the estimation of fewer parameters than other methods which assume a more complex covariance structure. This reduces the risk of overfitting the data, particularly if the variable dimensionality is relatively high. We conclude that the independence model may be a useful technique for building credit scoring models,

particularly in view of its conceptual and computational simplicity. It does, however, suffer from the disadvantage of estimating $P(g|\mathbf{x})$ indirectly.

Models which reduce the complexity of $P(g|\mathbf{x})$ by making less extreme independence assumptions can also be considered. Hand (1992) describes "*structured conditional probability distributions*" which use substantive knowledge about the relationships between characteristics to make independence assumptions.

4.4 Discrete multivariate techniques

Most of the standard techniques for building scorecards, such as discriminant analysis, assume continuous feature variables. However, the characteristics available for building credit scoring models are largely categorical (see Section 2.2). Therefore, although standard credit scoring methods can provide good classification accuracy, they may be sub-optimal. This motivates a review of multivariate methods for categorical data.

An important comparison study of discrimination techniques suitable for categorical feature variables is provided by Titterington et al. (1981). They describe the results of an experiment in which a range of parametric and non-parametric techniques are applied to the problem of diagnosing head-injured patients. Their study provides the focus for this review because they use a large complex data set that is similar in structure to typical credit data sets. In particular, the similarities are multidimensionality, a mixture of categorical and continuous data and a significant proportion of missing values.

4.4.1 The full multinomial model

The opposite extreme to the independence model of the last section is to make no assumptions about the data structure. This gives the full multinomial model. If there are p characteristics, each with a_i possible attributes, then the multinomial model estimates a separate class conditional probability estimate for each of the $\prod_{i=1}^{p} a_i$ possible cells, given by:

$$\hat{p}_{kj} = n_{kj} / n_j ,$$

where n_{kj} is the number of applicants from class j falling in cell k and n_j is the total number of applicants from class j. The multinomial model suffers from two principal disadvantages: first, the number of cells grows exponentially with the number of characteristics; this results in the data being too sparsely distributed for reliable probability estimates to be estimated for even moderate numbers of characteristics. Secondly, because the number of parameters to be estimated is relatively high, there is a danger of overfitting the data. Therefore, the multinomial model is not a practical method for building credit scoring models.

4.4.2 Lancaster models

Lancaster models (see Goldstein and Dillon, 1978) permit a whole range of structures to be fitted to the data ranging from the basic independence model to the full multinomial model. A common version of the method is to fit only first-order interactions between characteristics. Titterington et al. (1981) use this type of Lancaster model in their comparisons. The two-way marginal estimates for the model adopted are given by:

$$p(x_r, x_s|i) = \frac{n_i(x_r, x_s) + 1/(C_r C_s)}{N_i(r,s) + 1} \quad \text{for each } i,\ r,\ s,$$

where $n_i(x_r, x_s)$ and $N_i(r,s)$ are analogous to the corresponding terms defined for the independence model.

The authors describe two problems which arise with this type of model: first, negative cell estimates can be obtained (for cells where this occurred the independence model was used instead). Second, the numbers of cells may be too large when all the first order interactions are included, making it difficult to obtain reliable parameter estimates (one way to reduce this problem is to combine attributes for some characteristcs).

The comparison results were disappointing with the independence model outperforming the Lancaster models for three out of four datasets. We conclude that Lancaster models do not show potential for application to credit scoring.

4.4.3 Latent class models

Latent class analysis (see Fielding (1977)) involves the assumption that the probability estimate for $P(\mathbf{x}|g)$ can be decomposed into a weighted linear combination of underlying probability functions. Using notation from Titterington et al. (1981) the estimate for class i is given by:

$$p(\mathbf{x}|i) = \sum_{j=1}^{L} w_{ij} p_j(\mathbf{x})$$

Here L denotes the number of "latent" classes and the $\{w_{ij}\}$ are mixing weights.

Titterington et al. (1981) adopt a latent class model whereby the probability mixture distributions, $p_j(\mathbf{x})$, are represented by independence models (this greatly simplifies the model construction). The EM algorithm (see Dempster et al., 1977) is then used to estimate the maximum likelihood estimates. The classification results are disappointing, with the latent class models being outperformed by at least one of the independence and Lancaster models on each of the four data sets.

4.4.4 Loglinear models

Loglinear models represent a general approach to modelling categorical data. (See, for example, Bishop et al. (1975) or Fienberg (1977)). The model takes the form of a linear relationship between the natural logarithm of the expected cell counts and a set of parameters representing the effect of individual variables and interactions between variables. The general loglinear model does not specify a dependent variable and so allows more flexible examination of the structure of the data. However, in the credit scoring context creditworthiness represents a natural dependent variable and so this flexibility offers no advantage.

Loglinear models are best suited to low dimensional data and become very calculation intensive when dimensionality becomes high (as in our application). For high dimensional data one would also have to limit the number of interaction terms included, possibly to first order, to prevent overfitting of the data. The results for the Lancaster models described in Section 4.4.2 indicate

that loglinear models may not give better performance than the independence model.

4.4.5 Other methods

Titterington et al. (1981) use three other multivariate techniques for estimating $P(\mathbf{x}|g)$ for discrete data:

(1) A non-parametric kernel function in a factorised form proposed by Aitchison and Aitken (1976). They consistently perform less well than the range of other classification techniques considered. The authors point out that the dimensionality of the problem results in the data being too sparsely spread to obtain accurate estimates of the class conditional probabilities. We consider kernel methods in more detail in Section 4.8.

(2) The linear logistic model given by:

$$\frac{p(g|\mathbf{x})}{p(b|\mathbf{x})} = \exp(\alpha_g + \beta_g'\mathbf{x})$$ where α_i and β_i are parameters to be estimated.

This model performs reasonably well, given a crude method of handling missing data.

(3) Models involving treating the characteristics as discretizations of continuous feature variables and assuming multivariate normal distributions. This results in linear and quadratic discriminant rules depending upon whether the class covariance matrices are assumed to be equal. In comparisons the linear discriminant rule performed surprisingly well, beating the quadratic rule and several of the theoretically "more appropriate" methods.

Other classification methods for categorical feature variables that were not evaluated in this study are the location method (Krzanowski, 1975), methods based upon orthogonal series (Goldstein and Dillon, 1978) and Hill's nearest neighbour method (see Section 8.2.3.4).

In conclusion, despite the range of classification techniques available for dealing with categorical feature variables, no technique seems to consistently outperform the independence model and the linear discriminant rule. This

result indicates that methods which may not be conceptually appropriate, can give robust performance. This helps to justify a standard credit scoring practice of converting the characteristics into pseudo-continuous form using the weights of evidence transformation and using methods such as linear regression (see Section 2.2). In the rest of this chapter we review classification methods for continuous as well as categorical feature variables.

4.5 Regression techniques

Two of the most popular credit scoring techniques are linear and logistic regression. They both use a linear function of the predictor variables (characteristics) to model $P(g|\mathbf{x})$ directly. Because of their widespread use in the credit industry, we devote Section 7.2 to making a detailed theoretical and empirical comparison of the suitability of the two methods for credit scoring. Other regression approaches are now considered.

4.5.1 Regression models for ordinal data

McCullagh (1980) introduces a class of regression models for dealing with ordinal response variables. Two methods are proposed, the proportional odds and the proportional hazards models, both of which have the same general form given by:

$$link\{\gamma_j(\mathbf{x})\} = \theta_j - \beta'\mathbf{x}$$

where $\gamma_j(\mathbf{x})$ is the cumulative probability that the ordered response takes a value less than or equal to j for a characteristic vector $\mathbf{x}$ and θ_j and β are parameters.

Models for dealing with ordinal response data could be useful in the credit scoring context, in order to treat creditworthiness as a multi-state problem (rather than using the normal two class good/bad definitions). However, Hand (1992) describes a simulation study by Campbell et al. (1991), which compared ordinal approaches with multinomial logistic and linear discriminant analysis and concluded that ordinal models confer no advantage when used for classification purposes.

4.5.2 Nonparametric regression

One disadvantage of parametric regression procedures (such as linear or logistic regression) is that they impose restrictive assumptions on the model form. This can prevent complex dependencies in the data from being identified. A wide range of nonparametric regression techniques have been proposed to try and take account of this. Among the most widely studied are the kernel method (see Section 4.8), the k-NN method (see Chapter 8) and spline smoothing (Stone, 1977). These methods use local averaging in the predictor space to estimate $P(g|\mathbf{x})$. However, because of the sparsity of data in high dimensional spaces, these techniques do not generally perform well with multidimensional data (this is not true for the k-NN method).

A second approach to nonparametric regression is to use a recursive partitioning procedure (see Section 4.9). The general approach is to split the predictor space into two regions and then fit a separate linear model to the points in each region. The same procedure can then be applied recursively to each of the regions obtained in this way. The advantage of these methods over the local averaging methods described above (in higher dimensions), is that the local regions are determined by the response variation. This means that the influence of sparse regions of the characteristic space on the probability estimates is reduced. However, the recursive partitioning methods do suffer from the problem that the size of the sample available for model construction is drastically cut at each split.

4.5.3 Projection pursuit regression

Friedman and Stuetzle (1981) introduce a new nonparametric approach to multiple regression, which attempts to overcome the problems identified in the last section. It provides a more flexible regression surface than standard linear regression, by iteratively fitting a sum of smoothed functions of linear combinations of the predictor variables.

The procedure for estimating the regression surface at a particular point is as follows: at each successive iteration, project the data onto a plane spanned by the response and a linear combination of the predictors, $\alpha^t\mathbf{x}$ (projection step).

The linear projection of the predictors is smoothed using local averaging. Then, the parameter vector, α, is selected to maximise the amount of previously unexplained variance in the response, that is explained by the smoothed function of the predictors (pursuit step). The model at each iteration is given by the sum of the smoothed functions that were subtracted from the response to give the unexplained variance, and it summarises the structure identified so far.

Projection pursuit regression is suitable for multidimensional data because it uses a successive refinement approach, which allows complex data structures to be modelled without reducing the size of the sample available at each iteration. Interactions between variables are taken into account through the smoothing process. In addition, the results of each iteration can be represented graphically, thus aiding interpretation of the model structure. For these reasons, projection pursuit regression appears to offer potential for building credit scoring models. It is included in a general comparison of credit scoring techniques in Section 7.3.

4.6 Generalized linear models

This represents a generalization of the standard regression approach. A generalized linear model or GLM (see for example McCullagh and Nelder, 1983 and Dobson, 1983) is a model which specifies a relationship between a linear combination of the predictor variables ($\beta'\mathbf{x}$) and the expected value μ of the response y, where β is a vector of parameters. Here the response y is assumed to have a distribution from the exponential family. The general relationship between the predictor variables and μ can be expressed as

$$g(\mu) = \beta'\mathbf{x},$$

where the function $g(\mu)$, which is assumed to be monotone and differentiable, is called the *link function*. Particular GLM's are obtained by specifying a link function and a probability distribution for y.

The parameters of the models, β, are estimated using the method of maximum likelihood. In general the likelihood equations are non-linear and they have to be solved numerically by iteration using a procedure such as the *Newton-Raphson method* or the *method of scoring*.

Table 4.1 shows examples of common GLM's with their corresponding link functions.

Model	Link function, $g(\mu)$
Linear	μ
Gamma	$1/\mu$
Logistic	$\log(\mu/(1-\mu))$
Poisson	$\log(\mu)$
Probit	$\Theta^{-1}(\mu)$
Extreme value	$\log(-\log(1-\mu))$

Table 4.1: Some common GLM's.

We now consider selection of an appropriate GLM when the data has a binary response, as in the problem considered in this thesis. In this case the expected value of the response, μ, is equal to $P(y = 1)$. Under repeated sampling (taking N such random variables) the sum of the responses has the binomial distribution, $b(N, \mu)$. Thus, the binomial distribution (or the normal approximation to the binomial distribution) is the most appropriate distribution and logistic regression is theoretically the most appropriate GLM in our problem. Logistic regression is very widely used for analysing consumer credit data and is described in more detail in Section 7.2.2. The Poisson model is included in our comparisons in Section 7.3 to represent other GLM's. It is also appropriate for categorical reponse data, in particular for modelling counts.

4.7 Kernel methods

A common nonparametric approach to density estimation and discrimination problems is the kernel method. Hand (1982) shows how kernel methods have developed from the simple histogram and provides a general review of the standard work in this area. The general form of the multivariate kernel density estimator for a point **y** in the feature space is (Terrell and Scott, 1992):

$$\hat{f}(\mathbf{y}) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{\mathbf{x}_i - \mathbf{y}}{h}\right)$$

where the $\mathbf{x}_i$ are points in the design sample (size n), $K:\Re^p \rightarrow \Re^1$ is a kernel function centred at 0 that integrates to 1, and h is a smoothing parameter that is assumed to tend to 0 as n goes to ∞.

One of the most important aspects of kernel density estimation is the selection of a suitable value for the smoothing parameter, h. Early papers, such as Rosenblatt (1956), Parzen (1962) and Epanechnikov (1969), used a fixed h throughout the feature space. Terrell and Scott (1992) consider possible ways of improving the multivariate density estimates by varying h according to position in the feature space. In particular, three proposed approaches to varying h are considered.

(1) The k-Nearest Neighbour density estimate, proposed by Loftsgarden and Queensberry (1965), given by:

$$\hat{f}(\mathbf{y}) = \frac{k}{nV_p h_k(\mathbf{y})^p}$$

where $h_k(\mathbf{y})$ is the Euclidean distance from $\mathbf{y}$ to the kth nearest sample point, and V_p is the volume of the unit sphere in $\Re^p$. (The k-Nearest Neighbour estimator can be expressed in the general form of the kernel estimator described above, if K is chosen to be a uniform density on the unit sphere in p-dimensional space.)

(2) The adaptive kernel estimate of Breiman, Meisel and Purcell (1977), given by:

$$\hat{f}(\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_i^p} K\left(\frac{\mathbf{x}_i - \mathbf{y}}{h_i}\right)$$

where h_i is the Euclidean distance from x_i to the kth nearest other sample point. This becomes asymptotically equivalent to choosing $h_i \propto f(x_i)^{-1/p}$. A popular version of this method was introduced by Abramson (1982) who proposed using $h_i \propto f(x_i)^{-1/2}$ for all p.

(3) A generalised version of a suggestion due to Tukey and Tukey (1981) called a balloon estimator and given by:

$$\hat{f}(\mathbf{y}) = \frac{1}{nh(\mathbf{y})^p}\sum_{i=1}^{n} K\left(\frac{\mathbf{x}_i - \mathbf{y}}{h(\mathbf{y})}\right).$$

The authors investigate the performance of the kernel method using the different approaches to selecting h. Performance is measured using the asymptotic mean squared error at a point and the integrated version of this function over the real line. An important conclusion from their work is that it is surprisingly difficult to improve upon the kernel method with fixed h. (However, as the dimensionality increases, they show that the multivariate balloon estimate becomes increasingly more asymptotically efficient than the fixed kernel method.) They assert that, because of the difficulty of successfully applying optimal adaptive kernel methods, higher order fixed kernel methods seem a worthwhile option.

The most interesting results, from the credit scoring perspective, relate to the k-Nearest Neighbour method. Table 4.2 shows the relative efficiencies (using the asymptotic mean integrated squared error, AMISE) of the multivariate k-Nearest-Neighbour estimates to the fixed kernel for multinormal densities.

Number of variables	Relative efficiency
1	0
2	0
3	0.483
4	0.866
5	1.146
15	1.545
100	1.491

Table 4.2: Asymptotic relative efficiency of the k-Nearest Neighbour method to the fixed kernel method for different feature space dimensionalities.

The table shows that the k-Nearest Neighbour method performs very badly in low dimensions, but gives relatively superior performance to the kernel method in dimensions greater than 5. In particular, the maximum relative efficiency is achieved for 15 variables, the correct order of characteristics for credit scoring models. We conclude that the k-Nearest Neighbour method is more appropriate for application to credit scoring than other kernel methods. In Chapter 8 we

propose a k-Nearest Neighbour classifier for credit scoring, using an original distance metric.

4.8 Decision trees and decision graphs

Decision trees are nonparametric classification rules which split the feature space into regions of similar class membership probabilities (Friedman (1979) and Breiman et al. (1984)). Because the method works sequentially, it is often referred to as "recursive partitioning". In Section 4.6.2 we examined the recursive partitioning algorithm in a regression context. Several reasons why the method may be appropriate for credit scoring have been identified: first, it is able to fit complex non-linear relationships, because it does not have a restrictive parametric form; secondly, recursive partitioning constructs local regions based upon the nature of the response variation and so does not suffer from data sparsity in high dimensions (as local-averaging techniques like kernel methods do); and, thirdly, the underlying decision process may be better represented by a sequential process than a simultaneous one (where all the characteristics are considered simultaneously in the model). As was discussed in Section 3.1.2, decision trees have been used extensively for building credit scoring models and have given good classification performance in comparison studies (Srinivasan and Kim (1987), Boyle et al. (1992)). For these reasons, we include several variants of the standard decision tree method in our comparisons of classifier performance in Chapter 8. We adopt an inference scheme for constructing decision trees based upon the Minimum Message Length Principle, as proposed by Wallace and Boulton (1968). The methodology we use is described below, including a discussion of fanned decision trees and decision graphs, an extension of decision trees.

4.8.1 Decision trees

A decision tree consists of a series of sequential nodes, which split subsets of the feature space into descendent subsets by partitioning the characteristic values, and leaves, which specify a predicted class (or give predicted probabilities of class membership). (We restrict attention to decision trees with binary splits at each node.) An applicant is classified by following the nodes

which correspond to the applicant's characteristic values to the resulting decision leaf.

The idea behind the decision tree is that each split gives descendent nodes which are purer in terms of class than the parent node. We consider two components of constructing decision trees:

(1) The selection of rules for splitting a node (growing phase):
To begin with the decision tree consists of one leaf, which contains all points in the sample. The growing phase involves repeatedly splitting a leaf of the decision tree (and replacing the original leaf with a node) such that the dependent leaves provide better class separation. At each split all the currently unused characteristics are considered and the best splitting point for each characteristic is determined. To do this the attribute values of each characteristic are reordered into order of ascending proportion good (or bad). Various splitting rules have been proposed. Friedman (1979) proposes the myopic splitting rule which involves minimising the expected loss if this particular split is the only split. Brieman et al. (1984) consider more complicated splitting rules, including ones that can look ahead a number of splitting steps before determining the best split at the original node.

In this thesis we adopt an encoding procedure for decision trees based upon the Minimum Message Length Principle (Wallace and Boulton, 1968). The MMLP is a criterion for assessing the quality of different models, which involves encoding a model and corresponding data as a binary string and selecting the model which has the minimum message length. A procedure for encoding decision trees is proposed by Wallace and Patrick (1993). We use splitting rules which involve minimising the message length of the sets of decision trees obtained from looking ahead 1 and 2 steps.

(2) The decision when to declare a node as terminal (pruning phase):
The tree is grown using a splitting rule until the number of sample points at a particular leaf is less than a specified threshold or the possible splits are not able to improve class separation. The resulting tree is likely to overfit the design sample used to grow it. As a result, the tree is pruned by replacing decision nodes, whose children are leaves, by leaves if this satisfies some criterion (such as reducing the message length). The leaves of the final tree

produced are designated as good or bad for classifying future applicants, depending upon the proportion of goods and bads in the design sample and the criterion for performance. Many authors (e.g. Brieman et al. (1984)) consider pruning to be the most important part of the construction of trees.

4.8.2 Extensions of decision trees: fanned trees and decision graphs

Oliver J.J. (1994) proposes using *fanned* decision trees as an alternative to pruning. It is assumed that a decision tree has been grown using a splitting rule (such as minimisation of the message length for an encoded tree). Then, to classify applicants who fall into a particular leaf of this tree, we calculate the fanned set for this leaf. The fanned set of trees for a leaf are generated by using every possible characteristic in turn to expand the tree by one level. An applicant in this leaf is assigned class membership probabilities by averaging the corresponding probabilities over the fanned set of trees. The author implements a fanned decision tree procedure using the Minimum Message Length Principle (MMLP) and presents comparisons with the standard pruning criterion. The MMLP has the advantage over maximum likelihood approaches that the message length of a tree can be used as a measure of goodness of fit and, thus, it can be used to generate weights for the tree avearging process. The comparison results for seven data sets showed that the fanned decision trees almost always did as well as the pruning method and sometimes significantly outperformed the standard decision trees. We use fanned trees constructed using the MMLP in Chapter 8.

Decision graphs are another extension of decision trees (Oliver J.J., 1992). A decision graph is similar to a decision tree in that it consists of decision nodes and leaves and the method of categorizing objects is the same. It differs from decision trees in that it may include *Joins,* where a Join is represented by two nodes having a common child. The advantage of this is to allow two subsets of the population that may have common features to be joined into one subset. Oliver proposes constructing decision graphs using the MMLP. The procedure includes a parameter, P_j, corresponding to the probability of making a join, which allows one to determine the separation of the nodes in the decision graph. If $P_j = 0$ then the decision graph becomes a decision tree. The message length of a decision graph (for different P_j) can be used as a metric to

determine whether a decision tree or a decision graph is more appropriate for a particular data set. We examined the performance of decision graphs for different values of the parameter P_j in the comparisons presented in Section 7.3.

4.9 Neural networks

Neural networks are computer models which use representations of human neural systems to solve complex problems (see McClelland and Rumelhart (1986) or Beale and Jackson (1990)). They consist of interconnected nodes, representing neurons, which are usually hierarchically structured in layers. Each node receives a weighted sum of inputs and computes an output value using a suitable transformation. The most common transformation is the sigmoid function given by $f(x)=1/(1+e^{-kx})$, where k is constant which controls the spread of the function. The output value is then passed on to other connected nodes. In this way the neural network emulates the processing of information by the brain.

By choosing a suitable network architecture one can model complex statistical relationships between variables. A standard model is the multilayer perceptron, which consists of three layers of nodes: an input layer, an output layer and a layer in between the two, referred to as the hidden layer. Information passed into the input nodes propogates through the hidden layers to give an output at the output nodes. The nodes in the different layers of the network are connnected by links with variable weights, which correspond to the parameters of a statistical model.

A neural network classifier for credit applicants can be constructed by training a suitable network using the design sample. The input nodes correspond to the applicants' characteristic vectors and the output nodes correspond to creditworthiness. Construction of the network involves estimating the weights between nodes using a learning algorithm. The learning algorithm for the multilayer perceptron is called the "generalised delta rule" or the "backpropogation rule" (Rumelhart, McClelland and Williams, 1986). It modifies the network weights in order to minimise the difference between the

expected and actual output values (corresponding to creditworthiness) for the design set. The error function is given by:

$$Error = \frac{1}{2}\sum_j (E_j - O_j)^2$$

where E_j is the expected output value for an applicant at node j and O_j is the corresponding observed output. The error function is minimised using a gradient descent on the error surface in the weights space. The weights are modified from the output layer backwards according to

$$\Delta w_{ij} = \alpha \delta_j A_i$$

where w_{ij} is the weight from node i to node j, α is a scalar parameter, δ_j is an error term for node j and A_i is the network value at node i. For output nodes, the error term δ_j is given by

$$\delta_j^1 = O_j(1-O_j)(E_j - O_j),$$

and for nodes in the hidden layer it is given by

$$\delta_j^2 = O_j(1-O_j)\sum_k \delta_k^1 w_{jk},$$

where the sum is over the nodes in the output layer. The design set are passed through the network until the change in weights becomes negligible. The scalar parameter, α, is used to adjust the rate of gradient descent. In selecting this parameter, one needs to balancing the conflicting objectives of minimising learning time and avoiding local minima on the error surface. The resulting network can be applied to future applicants by inputting their characteristic values into the input nodes to give predicted probabilities of class membership at the output nodes.

Several alternatives to the multilayer perceptron have been considered in the neural networks literature (for an introduction see Beale and Jackson, 1990). One approach is to use a sum of nonlinear functions, known as radial basis functions, to partition the characteristic space. Instead of using hyperplanes, defined by weighted sums of the form $\sum w_{ij}x_i$, the radial basis approach uses hyperellipsoids, of the form $\phi(\|x-y\|)$, where $\|...\|$ is a distance measure. The advantage of the radial basis approach is that, once the radial basis functions have been selected, the problem reduces to optimizing a set of linear equations (whereas the multilayer perceptron involves optimizing nonlinear functions of $\sum w_{ij}x_i$). The difficulty with this method is selecting appropriate radial basis functions to enable the data structure to be modelled. Standard choices for the radial basis function and distance metric are $\phi(r) = e^{-r^2}$ and

$\|x-y\| = \sum_i (x_i - y_i)^2$ respectively. If the structure of the inputs (the characteristic vectors) is known a priori, then this can be used to provide suitable radial basis functions.

Other approaches to constructing neural networks are Kohonen's self-organising networks and Hopfield networks. The former approach uses a technique known as vector quantisation to organise the network so as to identify the structure of the input vectors. The method is called self-organising because it does not use response vectors from the training set in order to construct the network. It serves a similar function to cluster analysis and is unlikely to be of use for classifying applicants for credit. The latter approach constructs a network within which every node is connected to every other node in both directions and the weights between two nodes are symmetrical. The Hopfield network takes in an input at all nodes and iterates until a minimum on the error surface is reached. As with the Kohonen networks there is not an explicit mechanism for incorporating a response vector into the network.

Neural networks, in particular the multilayer perceptron described above, have several advantages over standard parametric methods of building credit scoring systems, such as discriminant analysis: first, they do not require the assumption of potentially unrealistic distributional forms; secondly, the hidden layer allows complex nonlinear relationships and dependencies between variables to be identified (several hidden layers can be used to fit more complex model forms); thirdly, the parallel nature of the method may be more appropriate for complex high dimensional data structures than a serial approach and, finally, the output layer can contain any number of nodes, thus enabling the response, creditworthiness, to be multivalued. (It can be shown that a neural network without a hidden layer is equivalent to a discriminant analysis and linear regression model.) Neural networks also do not suffer from the disadvantage of data sparsity in high dimensions shared by local averaging methods, such as the kernel method.

There has been considerable interest in the application of neural networks to the credit scoring problem, as indicated by papers such as Bazley (1993). Beale and Jackson (1990) report the results of a comparison of discriminant analysis with a multilayer perceptron for credit scoring. The results predicted that the

network would increase profitability by 7%. Comparison experiments in other related fields, such as Yoon et al. (1993), have demonstrated that neural networks can significantly outperform conventional statistical classification methods. However, there is a need for further research to establish the most appropriate forms of neural network for the credit domain (including the number and size of the hidden layers) and to provide rigorous comparisons with existing credit scoring techniques.

We end by noting that one particular weakness of neural networks is that they do not provide a means of describing the relative contribution of different characteristics to the classification rule. This could result in overfitting of the data. There is also a danger that the network might be seen as a "black box" by credit managers who would be responsible for its implementation. This could result in a loss of managerial control over the credit granting process. Yoon et al. (1991) describes a method of interpreting the relative importance of different characteristics by linearly approximating the relative strength between each input and output node. This can be used to provide a conventional scorecard. Further research is needed into the interpretation of neural network models.

4.10 Genetic algorithms

The genetic algorithm was introduced as a technique for assessing credit applicants in Section 3.2. It is a classification method which uses the principles of natural selection from population genetics to search efficiently through large noisy solution spaces (Holland (1975), Goldberg (1989) and South et al. (1993)).

The method starts by generating an initial parent population, usually randomly. Each member of the population is represented by a (binary) string, which represents a chromosome. A mating pool is selected from the parent population by selecting points which give the best values of an objective function. A new population is produced by applying genetic operators (crossover and mutation) to the mating pool and combining this with a proportion of points from the initial population. The crossover operator randomly selects complementary parts of two strings to combine. This enables the construction of new

individuals from existing individual who perform well, so that the algorithm can search new parts of the solution space. The mutation operator randomly changes a part of any string with a probability inversely proportional to the size of the population. This helps to prevent the premature loss of genetic material and compensates for sampling error. By applying these genetic operators to the mating pool, the "high performance" chromosomes are propogated through the population. This process is repeated until some stopping criteria is reached, such as an upper bound on the number of iterations.

There are two important components to the successful application of the genetic algorithm: first, the selection of a suitable representation of the solution space, which enables the genetic operators to select the "best" members of the population; and, secondly, selection of a suitable objective function to evaluate the difference between a member of the population and the required optimum. Both of these issues are discussed in detail by South et al. (1993).

Fogarty and Ireson (1993/4) apply the genetic algorithm (an incremental version) to credit scoring by using the method to identify ranges of attribute values for particular characteristics, or combinations of characteristics, which are common to low or high risk applicants. Estimates of the class membership probabilities are obtained by applying Bayes rule. The objective function is the accuracy of predictions for the training set, obtained by subtracting true class from the predicted outcome. This is used to provide a fitness measure for selecting members of the population. The classification results showed that the genetic algorithm has the potential to outperform other parametric and nonparametric classification methods applied to credit scoring.

South et al. (1993) highlight three reasons why the genetic algorithm can be useful for classification problems, like credit scoring, where the characteristic space is multidimensional with complex, non-linear relationships between the response and the characteristics:

(1) Genetic algorithms search from a population of solutions, rather than a single solution point. By using the crossover operator the genetic algorithm is able to search new parts of the solution space. This reduces the chance of the method converging to local optima.

(2) The genetic algorithm uses available "scoring information" to assess the value of potential solutions, rather than "auxiliary knowledge" (such as derivatives).

(3) Genetic algorithms are probabilistic rather than deterministic transition rules. This helps to reduce the chance of convergence to local optima and alleviate the influence of noise.

We conclude that the genetic algorithm show clear potential for building credit scoring models. Future research is needed to identify optimum values of the genetic algorithm parameters (the crossover probability, mutation probability, population size, generation gap and the reproduction and replacement scheme) and alternative ways of representing the data.

4.11 *n*th order Markov chain models

In Section 3.2.1 we discussed how Markov chain models have been used for modelling evolving credit portfolio performance over time (for example by Cyert et al., 1962). In this section we describe the methodology and show how it can be applied to the initial credit granting decision in order to make accept/reject decisions on individual applicants based on their characteristic vectors.

It is assumed that each month an account can belong to one of s states:

{paid, one month delinquent, ... , (s-1) months delinquent}.

The probability that an account in one state in one period will move into a particular state in the next period is defined by a transition matrix of probabilities. If $D(t,\mathbf{x})$ is the state after t months ($t>0$), given a characteristic vector $\mathbf{x}$, then the *1*st order Markov chain model involves estimating the transition probabilities:

$$P_{ij}(t,\mathbf{x}) = P(D(t,\mathbf{x}) = j | D(t-1,\mathbf{x}) = i).$$

The *n*th order Markov chain is defined in a similar fashion with conditioning on $D(t\text{-}2,\mathbf{x})$, $D(t\text{-}3,\mathbf{x})$, ..., $D(t\text{-}n,\mathbf{x})$. In other words, increasing the order of the Markov chain increases the dependency of future predictions on previous states.

Three approaches to modelling the $P_{ij}(t,\mathbf{x})$ in the *1*st order case are described below:

(1) *Nonstationary Markov chain*:

$$P_{ij}(t,\mathbf{x}) = g(f_{ij}(t,\mathbf{x})),$$

where $g(\bullet)$ is a link function (as used in *generalised linear models*) and

$f_{ij}(t,\mathbf{x}) = \beta_{1t}x_1 + \ldots \beta_{pt}x_p$ a linear function of the characteristics.

(2) *Stationary Markov chain*:

$$P_{ij}(t,\mathbf{x}) = g(f_{ij}(\mathbf{x}))$$

where $g(\bullet)$ and $f_{ij}(\mathbf{x})$ are defined as above.

(3) *Stationary Mover-Stayer model*:

$$P_{ij}(t,\mathbf{x}) = S_j(\mathbf{x}) + (1 - S_i(\mathbf{x}))P_{ij}(\mathbf{x})$$

where S_i is the proportion of *stayers* in state i. The "best" applicants in our problem could regarded as *stayers* because they have a very low probability of missing any repayments.

The parameters β_i are estimated using maximum likelihood estimation. The resulting model can be used to estimate the probability of an applicant being in a particular state at a particular point in the future. The advantage of this approach is its flexibility, allowing applicants' estimated creditworthiness to evolve over time. Research is needed to identify both the appropriate order of the Markov chain and the appropriate model assumptions.

PART 2

Fundamental aspects of credit scoring methodology

Chapter 5

Assessment of performance

5.1 Introduction

Assessment of classifier performance is an essential part of the construction of credit scoring models. The aim of the credit grantor should be to select performance measures which reflect the underlying objectives of the system. In the particular problems we investigate in this thesis the primary objective is identification of potential bad debt and so the most appropriate performance measure is the expected proportion of bad applicants in the accept region. (This criterion is discussed in Section 5.2.2.) However, other aspects of the credit granting decision may require different ways of assessing performance. This motivates a general review of approaches to measuring performance.

We distinguish between two broad approaches: the first involves assessing whether the classifier is "good enough" by some objective standard (*absolute performance*) and the second involves assessing whether the classifier is "better than" an alternative classifier (*relative performance*). Assessment of relative performance often involves using significance tests to compare absolute performance measures (see Section 5.3). In this way the two approaches can be complementary.

Measures of absolute performance can be further subdivided into measures of *discriminability* and *reliability* (Habbema et al. (1978), Hilden et al. (1978a) and Hand (1994)). Discriminability is concerned with measuring how successful a classifier is in assigning an applicant to the true class. A good classifier, in terms of discriminability, is one which provides high estimates of $P(c|\mathbf{x})$, where c is the true class, for applicants $\mathbf{x}$. Reliability, on the other hand, is concerned with measuring how accurately the classifier is estimating the true class membership probabilities $P(g|\mathbf{x})$.

The most widely used measure of discriminability (and performance in general) is probably the error rate. For this reason we make it the starting point for our discussion of performance measures (see Section 5.2.1). A simple estimate of the error rate is obtained by counting the proportion of misclassified applicants in the design set. However, this can lead to underestimates of the error rate for future cases sampled from the same distribution. We discuss various other approaches to error rate estimation including the hold-out method, crossvalidation, bootstrap methods and the average conditional error rate approach. In Section 5.3.2 we present a likelihood ratio test to compare the error rates from two classifiers.

The criterion for performance used in most of this thesis is the bad rate amongst the accepts, given a fixed acceptance rate (see Section 5.2.2). The bad rate is a discriminability measure which is often used in credit scoring problems and can be considered as a component of the error rate. The assumption of a fixed acceptance rate is unusual for a classification problem and has some implications for the assessment of performance. In particular, by fixing the acceptance rate we impose bounds on the best and worst performance that can be achieved. This issue is also discussed with relation to the error rate and the good rate amongst the rejects.

In Section 5.2.3 we present an overview of more sophisticated measures of discriminability (*continuous performance measures*) which assess the agreement between the estimated $P(g|\mathbf{x})$ and the actual outcomes. This is in contrast to the error rate and bad rate which use categorizations of the estimated $P(g|\mathbf{x})$. The advantages of using a continuous measure are assessed.

In some cases the accuracy of the estimated $P(g|\mathbf{x})$ may be more important than discrimination between classes. One example is the when the objective is to obtain accurate estimates of the true status of the rejects. In these situations reliability measures are appropriate. We describe several such measures in Section 5.2.4 and derive general statistics based upon these measures, which can be used to test the hypothesis that a classification rule is reliable. Some important weaknesses of the method are highlighted (see also Hand (1994)).

In Section 5.3 we turn our attention to assessing relative performance of classifiers. We have developed two tests for comparing the bad rates of

different classifiers. First, we present a significance test based upon Fisher's exact test. This involves removing applicants classified as accepts under two classifiers and comparing the remaining proportion of bads under each classifier. Secondly, we derive a likelihood ratio test which can be used to compare the bad rate. Both tests can be adapted to compare the error rate under two classifiers (as well as more general performance criteria). These two tests represent the major original contribution of this chapter.

5.2 Absolute performance

5.2.1 Error rate

The most common measure of classification performance is the *error rate*. Hand (1986) distinguishes between three types of error rate:

(1) The *Bayes error rate*, $e_B = \int [1 - \max P(i|\mathbf{x})] f(\mathbf{x}) d\mathbf{x}$, where $P(i|\mathbf{x})$ is the probability that an applicant with characteristic vector $\mathbf{x}$ belongs to class i (= g/b) and $f(\mathbf{x})$ is the overall mixture distribution. This is the minimum possible error rate given a set of characteristics.

(2) The true error rate: $e_T = \int_{\Omega_a} P(b)P(\mathbf{x}|b)d\mathbf{x} + \int_{\Omega_r} P(g)P(\mathbf{x}|g)d\mathbf{x}$, where Ω_a and Ω_r are the accept and reject regions respectively. This is the expected probability of misclassifying a new applicant.

(3) The expected error rate: this is the expected value of e_T over design sets of a particular size.

5.2.1.1 Traditional methods of error rate estimation

In the credit scoring context we are most interested in performance on future applicants and so the true error rate is most appropriate. If a test sample is available which is independent of the design sample used to construct the classifier then estimation of the error rate is simple. An unbiased estimate is

obtained by counting the proportion of incorrectly classified applicants in the accept and reject regions.

However, if the test data is available before the construction of the classifier, one might wish to add this to the design set (in order to use all the data to produce the best possible classifier). There has been much recent research into the identification of suitable methods of estimating the error rate in this situation (Toussaint (1974), Hand (1986)). (The approaches described below apply equally to estimation of the bad rate and other performance measures.)

One obvious approach when no test set is available is to build a classifier using the full sample and then reapply it to the full sample to obtain the proportion misclassified. This produces the *resubstitution* (or *apparent*) estimator which is likely to yield optimistically biased results (it underestimates the proportion of future cases that will be misclassified). As the design set size, N, increases this optimistic bias will decline.

A standard approach to error rate estimation, which is popular in the credit industry, is the *hold-out* method. This involves splitting the full data into two sub-samples, designing a classifier using one (the *design* set) and testing using the other (the *test* set). It has the advantage of computational simplicity. However, for small samples, it suffers from the disadvantage of not using all of the available data to design the classifier. The hold-out method was discussed in Section 3.1.1. It was found to provide reliable model coefficients and robust error estimates for different hold-out percentages. We concluded that when the sample size is large enough, it can be a suitable testing procedure. This is the standard approach we use in this thesis.

The hold-out method is a special case of a general approach to error rate estimation involving splitting the data into two sets, designing a classifier using one and testing using the other, then repeating the process with different splits and finally averaging the error rate results over the test sets. In Chapter 8 we adopt an extension of the hold-out method descibed by Leonard (1993). Other methods using this principle are the *rotation method* and the *leave-one-out method* (also called the *Lachenbruch method* or *cross-validation*).

The leave-one-out method involves taking a test set of a single point at each stage and using the rest of the sample (N-1 points) to build the classifier, with N splits being conducted. It was perhaps the most popular of the early methods because the bias due to using a reduced design set is small. However, later work has criticised it for having a relatively large variance and requiring extensive computing time.

Over recent years whole new classes of error rate estimators have emerged including the *bootstrap estimator* and the *average conditional error rate.* We describe these two classes in more detail.

5.2.1.2 Bootstrap methods

The bootstrap approach, proposed by Efron (1982, 1983), represents a major advance in the field of error rate estimation. It corrects the bias of the apparent error rate by estimating this bias from subsamples drawn from the original data. Let the original sample consist of N applicants for credit denoted by X and let F be the overall mixture distribution. If b is the expected optimism of the apparent error rate and e_A is the apparent error rate then the true error rate is estimated by:

$$e_T = b + e_A.$$

The bootstrap method estimates the nonparametric ML estimate of the bias term b from the observed empirical distribution, $\hat{F}$. In practice this estimate is generated by Monte Carlo methods. That is one draws random design samples $\tilde{X}$ from the empirical distribution and estimates the value of $(\tilde{e}_c - \tilde{e}_a)$ for each sample, where $\tilde{e}_c$ is the true error rate for the classifier designed using $\tilde{X}$ and using X as the true population and $\tilde{e}_a$ is the apparent error rate of the same classifier. The bias b is then estimated by taking the average (expectation) over the bootstrap samples to give:

$$\hat{b} = E_{\tilde{X}}(\tilde{e}_c - \tilde{e}_a).$$

Efron (1982) shows the similarity between the bootstrap method and the *jackknife*, a general statistical technique for removing bias. Efron (1983) introduces several variants on the standard bootstrap procedure: (a) the *randomised bootstrap* which is designed to compensate for the discontinuities in

the empirical distribution $\hat{F}$ when the true distribution F is smooth; (b) the *double bootstrap* which uses two applications of the procedure outlined above to help reduce the negative bias of the ordinary bootstrap; (c) a novel approach, referred to as the *632 estimator*, which uses a linear combination of estimators. A simulation experiment to describe the different estimators is described. Despite having the weakest theoretical foundations, the 632 estimator performed best.

The bootstrap method has received widespread attention in the statistical literature and several futher extensions of the ordinary bootstrap have been proposed (see for example Hand (1986)). Some authors, such as Srinivasan and Kim (1987) and Leonard (1988), have used the bootstrap method to reduce the resubstitution bias when constructing credit scoring models. However, we suggest two reasons why the bootstrap approach may be less suitable than the hold-out method in the problem considered in this thesis:

(1) The bootstrap method is most useful in problems where the sample sizes are small. (This is rare in the credit scoring context.)

(2) The results of simulations by Chernick et al. (1985) showed that the hold-out method (and their MC estimator) gave lower mean squared error rates than a range of bootstrap estimators when the true error rate was in the range 25-50%. This is the case for the classifiers we consider in Chapters 7 and 8.

5.2.1.3 Average conditional error rate methods

This approach, described by Hand (1986), involves splitting the true error rate e_T into two components using:

$$e_T = \int e(\mathbf{x}) f(\mathbf{x}) d\mathbf{x},$$

where $e(\mathbf{x})$ is the conditional probability of error given $\mathbf{x}$ and $f(\mathbf{x})$ is the corresponding overall mixture distribution. The probability functions $e(\mathbf{x})$ and $f(\mathbf{x})$ can then be estimated separately to give:

$$\hat{e}_T = \int \hat{e}(\mathbf{x}) \hat{f}(\mathbf{x}) d\mathbf{x}.$$

The advantage of this approach is that the mixture distribution $f(\mathbf{x})$ can be estimated entirely from unclassified data. Thus, in the credit context, rejected

applicants for credit with unknown true status can be used to improve the estimate of the true error rate.

To illustrate how the general approach works we describe an example presented by Hand (1986). Given an independent test sample of size N, let the conditional probability of error be given by the indicator function:

$$\hat{e}(\mathbf{x}_i) = I(c_i, \hat{c}_i),$$

where $\mathbf{x}_i$ is the characteristic vector for the ith point in the test set and c_i and $\hat{c}_i$ are the true and estimated classes for the ith point respectively. The overall mixture function for the test set is given by:

$$f(\mathbf{x}) = \begin{cases} 1/N & \mathbf{x} = \mathbf{x}_i \\ 0 & elsewhere \end{cases}.$$

Then the average conditional estimate of the error rate is given by:

$$\hat{e}_T = \sum_{i=1}^{N} I(c_i, \hat{c}_i).\frac{1}{N}.$$

Hand (1986) gives a general review of different approaches to average conditional error rate estimation taken over the last twenty years. Two sources of bias that can result from these methods are described: first, estimating the error rate at a point from a finite sample and secondly, resubstitution bias if the same data is used to design the classifier and estimate the conditional error rate.

5.2.2 Bad rate and our criterion for performance

The criterion for performance used in this thesis is rather unusual and deserves some discussion. Whereas most applications of classification procedures work with the error rate, for commercial reasons in this application the proportion to be accepted is pre-specified and the aim is to minimise the number of bad applicants accepted. This criterion is thus involved with only part of the misclassification space. This has some interesting consequences for the classification rule.

One immediate consequence is that it imposes bounds on the best and worst performance that can be achieved. Suppose that a proportion a of applicants must be accepted, and that the population contains a proportion p of good risk applicants. The *best* bound for the bad rate amongst the accepts is given by:

$$B_{br}(a,p)=\begin{cases}1-(p/a) & a>p\\ 0 & a\le p\end{cases}$$

It is obtained by accepting all good applicants and making up the rest of the $100p\%$ accepts from bads (note that it does not depend upon the number of applicants in the sample). The *worst* bound for the bad rate amongst the accepts is given by:

$$W_{br}(a,p)=\begin{cases}(1-p)/a & a>1-p\\ 1 & a\le 1-p\end{cases}$$

It comes from accepting all the bad applicants and making up the rest of the accepts from goods.

Corresponding bounds on the best and worst good rates for the rejects and error rates can also be found.

$$B_{gr}(a,p)=\begin{cases}0 & a>p\\ (p-a)/(1-a) & a\le p\end{cases}$$

$$W_{gr}(a,p)=\begin{cases}1 & a>1-p\\ p/(1-a) & a\le 1-p\end{cases}$$

$$B_{er}(a,p)=\begin{cases}(a-p) & a>p\\ (p-a) & a\le p\end{cases}$$

$$W_{er}(a,p)=\begin{cases}2-(a+p) & a>1-p\\ (a+p) & a\le 1-p\end{cases}$$

As an example, we consider the test sample used to assess the *k*-Nearest Neighbour method described in Table 8.5.1, which has p = 45.3%. Using the above expressions yields Figures 5.1, 5.2 and 5.3, showing, respectively, the best and worst bounds on the bad rate amongst accepts, the good rate amongst rejects, and the error rate. When a = 0.70, as was typically the case in the problems we considered, this leads to the results in Table 5.1.

Best	35.3%
Random	54.7%
Worst	78.1%

Table 5.1: Bounds on the bad rate for a 70% acceptance rate.

Figure 5.1: Best and worst bounds on the bad rate amongst the accepts

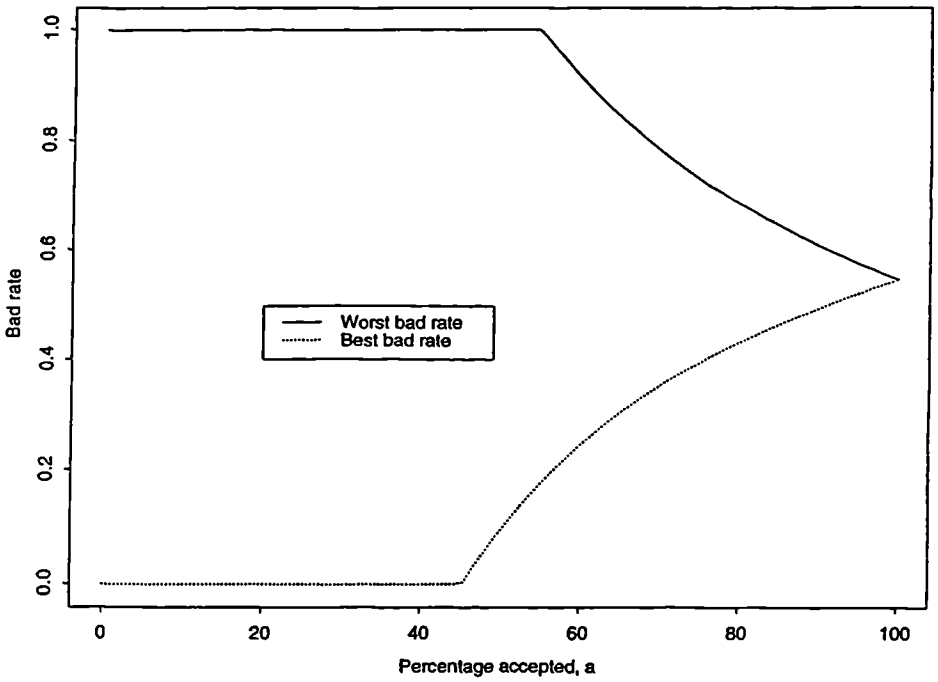


Figure 5.2: Best and worst bounds on the good rate amongst the rejects

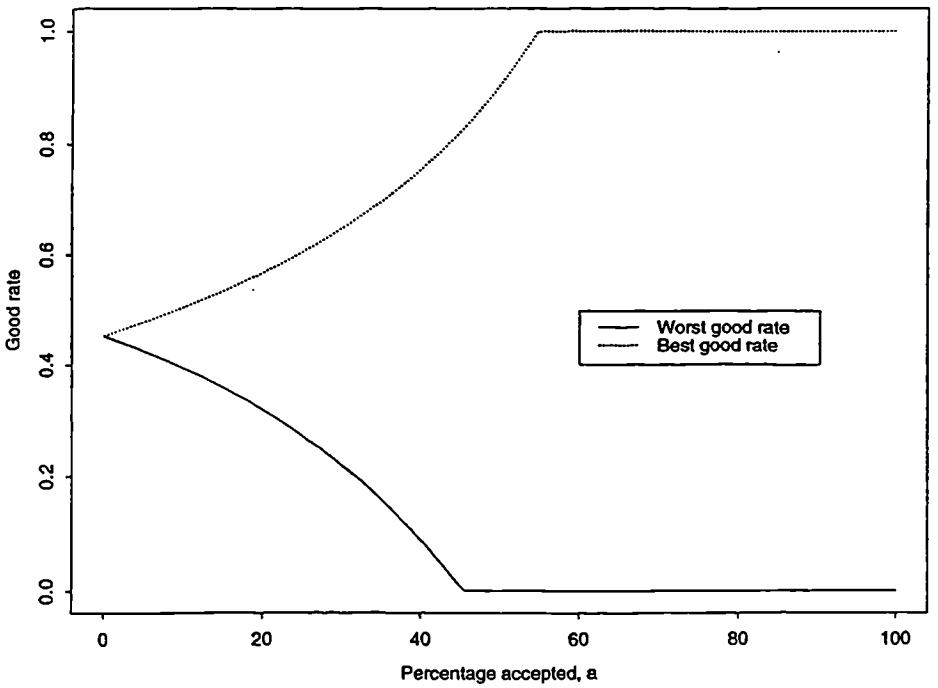
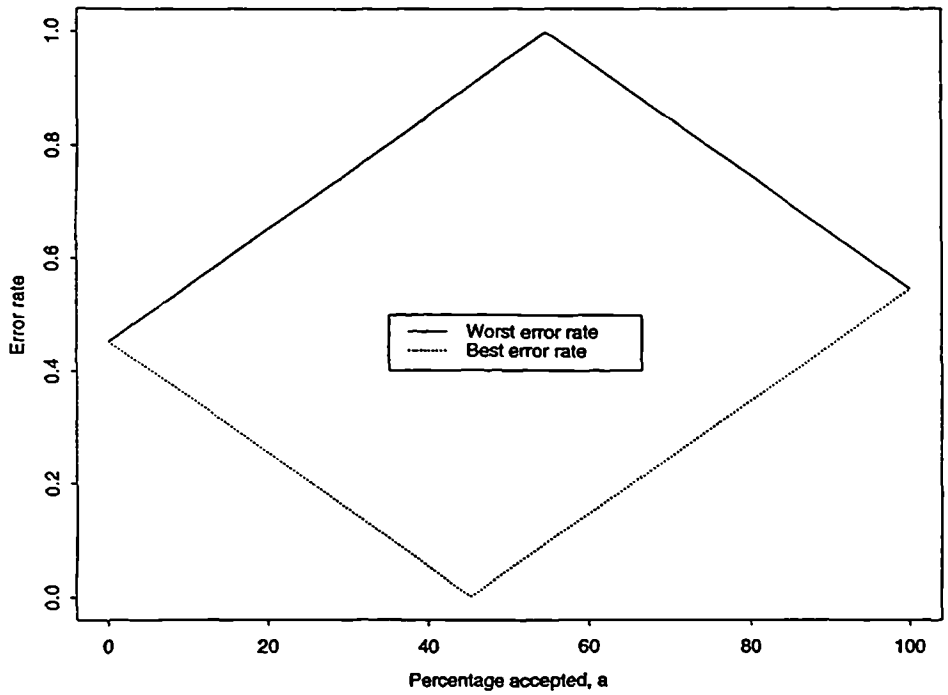


Figure 5.3: Best and worst bounds on the error rate



These results allow us to put the performance of the classification methods we consider into perspective. For example, while in general a bad rate of 36% might represent disappointingly poor performance, in this context it is only just above the theoretical best.

Note that it is the specification of a fixed acceptance rate which imposes these bounds. In other applications a common classification rule is to assign a case to the class i for which the estimated $P(i|\mathbf{x})$ is highest (or $\pi_i P(i|\mathbf{x})$ where π_i is the population prior for class i). This allows the acceptance rate to vary with the sample. In this situation, if the classes are perfectly separable, it is theoretically possible for a classifier to achieve zero error rate, whereas in our problem, even with perfect separability, this may not be the case.

5.2.3 Measures of Discriminability

In the introduction to this chapter we defined a discriminability measure to be one which measures how successful a classifier is in assigning an applicant to the true class. A good classifier, in terms of discriminability, is one which provides high estimates of $P(c|\mathbf{x})$, where c is the true class, for applicants $\mathbf{x}$. The bad rate and the error rate are specific examples of this general class of performance measures. In this section we give a review of other discriminability measures and evaluate their potential for assessing credit scoring models. We distinguish between measures based on counts of misclassifications, such as the error rate and bad rate, and measures which assess the agreement between the estimated $P(g|\mathbf{x})$ and the actual outcomes (continuous performance measures). We finish by considering the idea of a *strictly proper* scoring rule which can be used to select between alternative discriminatory measures.

5.2.3.1 Measures based on counts of misclassifications

One of the weaknesses of the error rate is that it gives no indication of which type of misclassification accounted for an error (classifying a good as a reject or a bad as an accept). In order to define other measures based upon the misclassification counts we consider the general 2 by 2 classification table shown in Table 5.2.

	Good	Bad
Accept	a	b
Reject	c	d

Table 5.2: Classification table resulting from a credit scoring model.

The range of common performance measures are described below:

(1) The error rate is given by $(b + c)/(a + b + c + d)$ and the non-error rate is given by $(a + d)/(a + b + c + d)$

(2) The bad rate can be expressed as $b/(a + b)$. In other contexts it is more common to consider the positive predicted value $a/(a + b)$ and the negative predicted value $d/(c + d)$.

(3) In epidemiology the terms *false positive* and *false negative* are used to refer to the rates $b/(b + d)$ and $c/(a + c)$ respectively.

(4) The complements of the false positive and the false negative are the sensitivity and the specificity given by $a/(a + c)$ and $d/(b + d)$ respectively.

(5) The prevalence (the prior probability of being a good risk) is given by $(a + c)/(a + b + c + d)$.

Although these measures are all interrelated the choice of which measure to use in a particular problem can be important. Hand (1994) gives a medical example to demonstrate how the sensitivity and specificity measures can contradict the positive predictive value.

The class of measures considered above provide a simple but useful description of the performance of a classifier. However, they suffer from the limitation of using categorizations of the estimated $P(\mathrm{g}|\mathbf{x})$ rather than the actual values themselves. This may not be important in the credit scoring context where the objective is to divide the applicant population into two categories (accept/reject)

in order to optimise a criterion such as the proportion of potential bad debt accepted.

5.2.3.2 Continuous measures of performance

In this section we consider the class of continuous performance measures which assess the agreement between the estimated $P(g|\mathbf{x})$ and the actual outcomes (Hilden et al. (1978b)). The measures that we consider are based on an average over applicants of a function of the estimated $\hat{P}(c_i|\mathbf{x}_i)$, where c_i is the true class of the ith applicant and $\mathbf{x}_i$ is the corresponding characteristic vector. The most important measures are considered in turn:

(1) The average probability assigned to the good/bad classes actually present:

$$C_1 = \frac{1}{N}\sum_{i=1}^{N}\hat{P}(c_i|\mathbf{x}_i), \quad [0, 1, \text{high}].$$

The minimum, maximum and desirable values are shown in sharp brackets after the formula. This measure has an intuitive appeal because the higher the probability assigned to an applicant's true class the higher the score. This measure is used in the evaluation of our reject inference method 6 in Section 6.6.2.2.

(2) The quadratic score or Brier score:

$$C_2 = \frac{1}{N}\sum_{i=1}^{N}\left\{\left(1-\hat{P}(c_i|\mathbf{x}_i)\right)^2 + \sum_{j\neq c_i}\hat{P}(j|\mathbf{x}_i)^2\right\} \quad [0, 2, \text{low}].$$

This measure has the particular advantage that not only does it take into account the predicted probability of an applicant's true class but also the size of the predicted probabilities of an applicant belonging to the other class(es). This only becomes valuable when the problem involves more than two classes.

The Brier score is intuitively reasonable because it takes low values (near to 0) when the predicted probabilities for all the applicants' true classes are high, and high values (near to 2) when these predicted probabilities are low.

(3) The logarithmic score:

$$C_3 = \frac{1}{N}\sum_{i=1}^{N}\ln\left(\hat{P}(c_i|\mathbf{x}_i)\right) \quad [-\infty, 0, \text{high}].$$

This measure has the undesirable property that if just one applicant has $\hat{P}(c_i|\mathbf{x}_i) = 0$ then C_3 becomes minus infinity. Therefore, it is usual to consider a modified statistic which penalises very small and zero values of $\hat{P}(c_i|\mathbf{x}_i)$ about equally.

One such modified logarithmic score is given by:

$$C_4 = \frac{1}{N}\sum_{i=1}^{N} \ln\left(\hat{P}(c_i|\mathbf{x}_i) + \varepsilon\right) \quad [\ln(\varepsilon),\ \ln(1+\varepsilon),\ \text{high}],$$

where ε is an arbitrary constant (0.001 is sometimes used). In practice the Brier score C_2 may be preferable to C_4 because it does not require inclusion of the constant ε.

We identify three potential advantages of this class of measures over the measures described in Section 5.2.3.1 (see Shapiro (1977) for further details):

(1) The principle disadvantage of the error rate and bad rate is that they are insensitive. In other words no account is taken of the magnitude of an error. For example, a bad applicant scoring just above the acceptance threshold counts the same as a bad applicant scoring the maximum possible number of points. In the credit scoring problem it is sometimes desirable to assess the accuracy of predictions near to the threshold. In this situation it is necessary for the performance measure to take account of the difference between an applicant's score and the threshold. Continuous measures such as the Brier score are able to do this.

(2) Continuous measures allow one to distinguish a decision from the evidence on which it is based. When the error rate is used the only information available for evaluation of the classifier is whether the prediction was true or false. This factor seems to be more applicable to the medical domain than the credit domain.

(3) In some cases continuous performance measures can reduce the variance of an estimator. To see this we make a comparison of the non-error rate (*NER*) described in Section 5.2.3.1 with the continuous measure C_1.

We begin by expressing the *NER* in a different form:

$$NER = \frac{1}{N}\sum_{i=1}^{N} I(c_i, \hat{c}_i),$$

where c_i and $\hat{c}_i$ are the true and estimated classes of the ith applicant and $I(c_i,\hat{c}_i)$ is the appropriate indicator function. It is immediately clear that *NER* and C_1 are measuring the same thing and that $E(NER) = E(C_1)$. However, NER is the sum of a random sample of zeros and ones whereas C_1 is the sum of a random sample of values in [0,1] which are unlikely to be exactly zero or one. Hence the variance of C_1 will be less than the variance of *NER*.

This argument cannot be generalised to include all continuous performance measures because they each take a different form.

5.2.3.3 Proper and strictly proper measures of performance

A property that helps one to choose between performance measures is *properness* (Hilden et al. (1978b)). The measures that we have considered can all be expressed in the form

$$C = \frac{1}{N}\sum_{i=1}^{N} f[\hat{P}(c_i|\mathbf{x}_i)].$$

Such a measure is said to be strictly proper (SP) if it satisfies

$$\sum_{j=1}^{2} P(j|\mathbf{x}) f[P(j|\mathbf{x})] > \sum_{j=1}^{2} P(j|\mathbf{x}) f[\hat{P}(j|\mathbf{x})],$$

where the summation is over classes, j, and $P(j|\mathbf{x})$ is the true (unknown) probability that an applicant with characteristic vector $\mathbf{x}$ belongs to class j. A proper measure satisfies the above relation with $\geq$ instead of $>$.

If the above relation holds it means that the performance measure gives the highest expected score to a classifier which estimates the true $P(j|\mathbf{x})$ perfectly. SP measures also give higher scores the nearer the predicted $\hat{P}(j|\mathbf{x})$ are to the true values. These are clearly desirable properties of a performance measure. In fact, the SP criterion is similar to the reliability criterion discussed in Section 5.2.4.

The SP criterion can be used to decide between alternative measures of performance. Of the measures we have considered, C_2 and C_3 are strictly proper, C_4 is approximately proper, the NER is non-strictly proper and C_1 is non-proper.

5.2.4 Reliability

Reliability measures (Hilden et al., 1978a) assess the accuracy of the predicted class membership probabilities $\hat{P}(g|\mathbf{x})$. We give two simple examples to illustrate how reliability differs from discriminability:

(1) Let $\hat{P}(g|\mathbf{x}) = P(g|\mathbf{x}) = k$, where k is a constant, for all characteristic vectors $\mathbf{x}$. Although the estimated class membership probabilities are equal to the corresponding true probabilities, the classifier is not able to distinguish between the good and bad classes. Therefore, the classifier is reliable but non-discriminating.

(2) Let $\hat{P}(g|\mathbf{x}) > t \Leftrightarrow P(g|\mathbf{x}) > t$ (for a threshold t) and $\hat{P}(g|\mathbf{x}) \approx \begin{cases} 0 \\ 1 \end{cases}$ for almost all characteristic vectors $\mathbf{x}$. This implies that the classifier achieves good discriminability. Now if we add the condition that $\hat{P}(g|\mathbf{x})$ is far from $P(g|\mathbf{x})$ $\forall \mathbf{x}$ and subject still to the above, then the classifier is unreliable.

Unfortunately, as noted above, the true $P(g|\mathbf{x})$ are unknown and so we cannot make a direct comparison of $\hat{P}(g|\mathbf{x})$ and $P(g|\mathbf{x})$. Hilden et al. propose a general procedure for deriving reliability measures based upon discriminability measures.

Let the general discriminability measure be given by

$$D = \frac{1}{N}\sum_{i=1}^{N} f(\hat{P}(c_i|\mathbf{x}_i)) \qquad (5.2.1),$$

where $f(\bullet)$ is a function which specifies the particular measure. To simplify the notation we write $p_i = \hat{P}(c_i|\mathbf{x}_i)$ in what follows. The general approach is to define a reliability measure as the deviation of a discriminability measure from its expectation, assuming it to be perfectly accurate. This last assumption is somewhat analogous to the null hypothesis assumption used in hypothesis testing. (We return to this analogy below.)

A general reliability measure can be written as:

$$R = \frac{1}{N}\sum f(p_i) - E^*\left(\frac{1}{N}\sum f(p_i)\right)$$

$$= \frac{1}{N}\sum[f(p_i) - E^*(f(p_i))], \text{ where } E^* \text{ denotes expectation}$$

with respect to the hypothesis of perfect reliability. The expectation can be expressed as

$$E^*(f(p_i)) = p_i f(p_i) - (1-p_i)f(1-p_i).$$

Thus,

$$R = \frac{1}{N}\sum[f(p_i) - p_i f(p_i) - (1-p_i)f(1-p_i)]$$

$$= \frac{1}{N}\sum(1-p_i)[f(p_i) - f(1-p_i)] \qquad (5.2.2).$$

We use the general form (5.2.2) to derive reliability measures for two specific discriminability measures:

(1) The average probability assigned to the true class, C_1.

In this case the function $f(p_i) = p_i$. Equation (5.2.2) reduces to

$$R_1 = \frac{1}{N}\sum(1-p_i)(2p_i - 1).$$

Figure 5.4 shows a plot of the contribution to R from one applicant against p.

(2) The non-error rate, $NER = \frac{1}{N}\sum_{i=1}^{N} I(c_i, \hat{c}_i)$.

Here $f(p_i) = I(c_i, \hat{c}_i)$ and equation (5.2.2) becomes

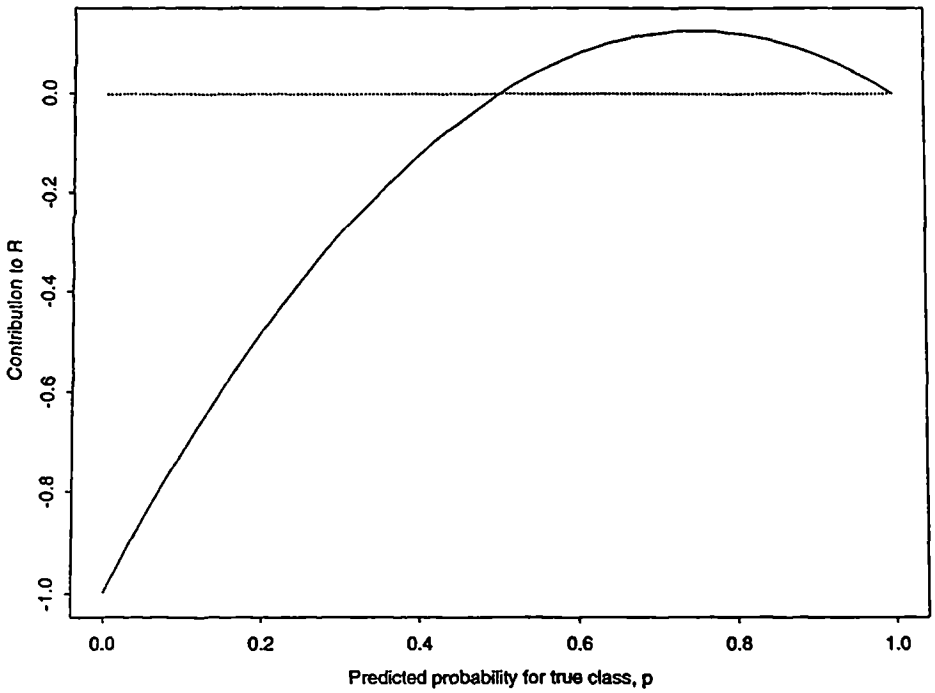
$$R_2 = \frac{1}{N}\sum(1-p_i)[I(c_i, \hat{c}_i) - I(\tilde{c}_i, \hat{c}_i)],$$

where $\tilde{c}_i$ is the complement of the true class for the ith applicant.

Hilden et al. derive reliability measures for the other performance measures considered in this chapter. The general procedure adopted for assessing the value of a reliability measure is to perform a significance test using the assumption of perfect reliability as the null hypothesis. Under the null hypothesis the expectation of R is zero. Thus, the appropriate test statistic is given by $T = \frac{R}{S}$ where S is the standard deviation of R. To calculate S we note that for a reliability measure R based on a discriminability measure D (in the general form 5.2.2):

$$Var^*(R) = E^*(R^2) - E^*(R)^2 = E^*(R^2) = Var^*(D).$$

Fig 5.4: The contribution to R from one applicant for different p



By independence of the observations we can write,

$$
\begin{aligned}
Var^*(D) &= \frac{1}{N^2}\sum Var(f(p_i)) \\
&= \frac{1}{N^2}\sum [E^*(f(p_i)^2) - E(f(p_i))^2] \\
&= \frac{1}{N^2}\sum [p_i f(p_i)^2 + (1-p_i)f(1-p_i)^2 - (p_i f(p_i) + (1-p_i)f(1-p_i))^2] \\
&= \frac{1}{N^2}\sum p_i(1-p_i)[f(p_i)^2 + f(1-p_i)^2 - 2f(p_i)f(1-p_i)] \\
&= \frac{1}{N^2}\sum p_i(1-p_i)[f(p_i) - f(1-p_i)]^2 \qquad (5.2.3).
\end{aligned}
$$

Combining (5.2.2) and (5.2.3) gives

$$
T = \frac{R}{\sqrt{Var(D)}} = \frac{\sum (1-p_i)[f(p_i) - f(1-p_i)]}{\sqrt{\sum p_i(1-p_i)[f(p_i) - f(1-p_i)]^2}} \qquad (5.2.4).
$$

Assuming normality of the test statistic, we can compare T with the appropriate normal critical values. Using a 5% significance value the null hypothesis of reliability is rejected if $|T| > 1.96$. Equation (5.2.4) can be used to derive test statistics for reliability measures based upon any of the discriminability measures considered in this chapter.

We discuss two limitations of the above approach to reliability:

(1) The general reliability measure R, given by (5.2.2), is equal to zero when $p_i = \begin{cases} 1/2 & \forall i \\ 1 & \forall i \end{cases}$ or $f(p_i)$ is a symmetric function, regardless of the true class membership probabilities. For example, this means that a classifier which assigns $\hat{P}(g|\mathbf{x}) = 1/2$ for all applicants will be considered perfectly reliable using any measure R.

The result of this property is that the reliability measure can take values indicative of a reliable classifier when, in fact, it is highly unreliable. To some extent this will be mitigated by the effect of the standard deviation on the test statistic T. However, in the extreme example where $\hat{P}(g|\mathbf{x}) = 1/2$ for all applicants, the standard deviation is zero resulting in an undefined value of T.

We note that in most practical situations the true $P(g|\mathbf{x})$ will range from 0 to 1. When this is the case the above approach can provide a useful measure of the reliability of a classifier.

(2) The significance test for reliability presented above asssumes that, under the hypothesis of perfect reliability, the reliability measure R is normally distributed. Given the shape of the curve in Figure 5.4 it might be more reasonable to expect the distribution to be skewed to the left.

In conclusion, despite limitations of the measures discussed above, reliability can be a useful guide to classifier performance. Hilden et al. (1978a) assert the importance of reliability in the medical context. As an example, it is important to be able to place trust in the estimated probability of a patient having a fatal disease. Reliability of a classifier is generally less important in the credit scoring problem where the objective is to discriminate between good and bad risks. However, it can play a useful secondary role by indicating whether a classifier's predictions are too optimistic/pessimistic and whether it favours either class. Furthermore, improving the reliability of a classifier often results in an involuntary improvement in discriminability. In Section 6.6.2.2 the reliability measure R_1 is used to assess the accuracy of reject inference method 6 in predicting the proportion of goods in the reject region.

5.3 Relative performance

In Section 5.2 we discussed measures of absolute performance. For commercial reasons, in this thesis, we choose to adopt the bad rate amongst the accepts, given a fixed acceptance percentage. Having constructed classifiers and calculated performance statistics, such as the bad rate, it is important to be able to assess whether any differences are due to chance or can be attributed to any real difference in performance. We have developed two tests for assessing these differences.

5.3.1 Significance test based on Fisher's exact test

5.3.1.1 Theory of the test

The general framework of this test can be adapted to compare a range of different performance measures. We begin by comparing the bad rates under two classifiers, S_1 and S_2. Both classifiers are applied to an independent test set to give a score for each applicant in the sample, which is then compared with a threshold leading to an accept/reject decision on that applicant.

Let n_{ijk} be the number of applicants from the test sample with the following properties:

$$i = \begin{cases} 1 & good \\ 2 & bad \end{cases}$$

$$j = \begin{cases} 1 & S_1\ accept \\ 2 & S_1\ reject \end{cases}$$

$$k = \begin{cases} 1 & S_2\ accept \\ 2 & S_2\ reject \end{cases}.$$

The above information can be expressed in tabular form as shown in Table 5.3.

		S_1 accept	S_1 reject
S_2 accept	Good	n_{111}	n_{121}
	Bad	n_{211}	n_{221}
S_2 reject	Good	n_{112}	n_{122}
	Bad	n_{212}	n_{222}

Table 5.3: A general 2 x 2 x 2 classification table for comparing classifiers S_1 and S_2.

The overall bad rates for S_1 and S_2 are $\frac{n_{211}+n_{212}}{n_{111}+n_{211}+n_{112}+n_{212}}$ and $\frac{n_{211}+n_{221}}{n_{111}+n_{211}+n_{121}+n_{221}}$ respectively. However, we cannot conduct straightforward tests on these two proportions because the observations are not independent. To conduct a test we need to eliminate the common elements

(since they are common this will not influence the conclusion about the relative performance of S_1 and S_2). This yields the 2 x 2 table shown in Table 5.4.

	Bads	Goods
$\overline{S}_1$	n_{212}	n_{112}
$\overline{S}_2$	n_{221}	n_{121}

Table 5.4: 2 x 2 classification matrix of applicants accepted using either S_1 or S_2 but not both.

The two reduced accept samples $\overline{S}_1$ and $\overline{S}_2$ are now independent. The rates we compare for S_1 and S_2 are $\frac{n_{212}}{n_{112}+n_{212}}$ and $\frac{n_{221}}{n_{121}+n_{221}}$. In other words we compare the bad rate amongst applicants accepted by S_1 but rejected by S_2 with the bad rate amongst those accepted by S_2 but rejected by S_1.

A comparison of the bad rates for the two classifiers S_1 and S_2 is equivalent to testing the null hypothesis of independence between rows in the 2 x 2 contingency table shown in Table 5.4. This is an example of a standard problem that has received considerable attention in the statistical literature over recent years (see, for example, Upton, 1982, Yates, 1984). The appropriate test is Fisher's exact test.

When the two margins of the 2 x 2 table are fixed the distribution of the cell counts is hypergeometric. The exact test generates the probability of obtaining the observed cell counts under the null hypothesis of row independence, given this distribution. The resulting probability estimate can be compared with a significance level leading to either rejection or no rejection of the null hypothesis. For more details of the exact test see Yates (1984) or Kendall and Stuart (1977). If the null hypothesis is rejected then we conclude that there is a statistically significant difference between the proportion bad amongst applicants accepted under one but not both of S_1 and S_2. We note that the proposed test is two-tailed.

There may be other quantities, apart from the bad rates, that one is interested in comparing, and this can be achieved using the general framework outlined above. As an example, we consider the overall proportions misclassified by S_1 and S_2 (the error rate). Using the above notation these proportions are

$\frac{n_{211}+n_{212}+n_{121}+n_{122}}{N}$ and $\frac{n_{211}+n_{221}+n_{112}+n_{122}}{N}$ respectively, where $N=\sum_{i,j,k} n_{ijk}$. As before, to test the hypothesis that the two proportions are equal we must eliminate the common elements. Thus, the two proportions that we need to compare are $\frac{n_{212}+n_{121}}{n_{112}+n_{212}+n_{121}+n_{221}}$ and $\frac{n_{221}+n_{112}}{n_{112}+n_{212}+n_{121}+n_{221}}$. Since these both have the same denominator, this is equivalent to testing the hypothesis that $\frac{n_{212}+n_{121}}{n_{112}+n_{212}+n_{121}+n_{221}}=\frac{1}{2}$. Fisher's exact test can be applied to the reduced samples as before.

5.3.1.2 A numerical example

To illustrate how the proposed significance test works in practice, we describe the results of comparing two reject inference methods (extrapolation and Method 5 from Section 6.6.1). In both cases a classifier was constructed using linear regression with eleven characteristics. An independent validation sample consisting of 2020 goods and 2285 bads was split into accepts and rejects using a 70% acceptance rate. The common elements in the two accept samples were eliminated to give the reduced samples shown in Table 5.5.

	bads	goods
Extrapolation	129	61
Method 6	122	85

Table 5.5: 2 x 2 classification table for independent accept samples.

Fisher's exact test was applied to the table giving a *p*-value for the one-sided critical region of 0.0404. Therefore, adopting a significance level of 5%, we reject the independence hypothesis, and conclude that the classifier from Method 5 gives improved discrimination amongst the reduced accept samples.

5.3.2 Likelihood ratio test

In this section we present an alternative test for comparing two classifiers, using a likelihood approach, which addresses a subtly different question from the test

considered above. Instead of comparing the values of performance measures for applicants accepted under one but not both of two classifiers, it compares the values of performance measures obtained from all applicants in the sample. In this sense it can be considered as a more powerful test of performance. As with the first test it can be adapted to consider the whole range of performance measures based on counts (described in Section 5.2.3.1). In particular, we focus on comparing the error rate and bad rate.

5.3.2.1 Comparison of error rates

We derive the likelihood ratio test for comparing the error rates when two classifiers S_1 and S_2 are applied to a test set which is independent of the original design set. Let n_{ijk} and p_{ijk} be the respective cell count and cell probability for the cell (i,j,k), where as before:

$$i = \begin{cases} 0 & bad \\ 1 & good \end{cases}$$

$$j = \begin{cases} 0 & S_1 \ reject \\ 1 & S_1 \ accept \end{cases}$$

$$k = \begin{cases} 0 & S_2 \ reject \\ 1 & S_2 \ accept \end{cases}.$$

The log-likelihood for the full data is given by:

$$L = \sum_{i,j,k} n_{ijk} \log(p_{ijk}) \qquad (5.3.1)$$

where $\sum_{i,j,k} p_{ijk} = 1$ and $\sum_{i,j,k} n_{ijk} = N$.

The likelihood ratio test involves comparison of the log-likelihood under two conditions:

(1) Assuming no relationship between the parameters.

(2) Assuming that the expected probability of an error is the same under both classifiers. This condition can be expressed as

$$P(g,r_1) + P(b,a_1) = P(g,r_2) + P(b,a_2),$$

where a_i and r_i are acceptance or rejection under classifier i. In terms of the p_{ijk} this condition can be re-expressed as:

$$p_{101} + p_{100} + p_{010} + p_{011} = p_{110} + p_{100} + p_{011} + p_{001},$$

which reduces to

$$p_{101} + p_{010} = p_{110} + p_{001} \qquad (5.3.2).$$

The condition (5.3.2) is the null hypothesis assumption of no difference in error rate.

We derive the ML parameter estimates separately for the two conditions:

Case (1): No constraints

Using the Lagrange multiplier, λ, define:

$$F = L + \lambda\left(\sum_{i,j,k} p_{ijk} - 1\right).$$

Then the likelihood equations are given by:

$$\frac{\partial F}{\partial p_{ijk}} = \frac{n_{ijk}}{p_{ijk}} + \lambda = 0$$

$$\therefore \hat{p}_{ijk} = -\frac{n_{ijk}}{\lambda},$$

Applying the condition $\sum_{i,j,k} p_{ijk} = 1$,

$$-\sum_{i,j,k} n_{ijk} = \lambda$$

$$\therefore \lambda = -N$$

$$\therefore \hat{p}_{ijk} = \frac{n_{ijk}}{N}, \quad \text{for } i,j,k \in \{0,1\}.$$

Case (2): Constrained

In this case we have to maximise the log-likelihood function subject to the conditions

(a) $\sum_{i,j,k} p_{ijk} = 1$

(b) $p_{101} + p_{010} = p_{110} + p_{001}$.

Again we use Lagrange multipliers:

$$G = L + \lambda_1(\sum_{i,j,k} p_{ijk} - 1) + \lambda_2(p_{101} + p_{010} - p_{110} - p_{001})$$

The likelihood equations take three distinct forms according to the p_{ijk}:

(i) $P=\{p_{111},p_{100},p_{011},p_{000}\}$:

Let p represent a general element from the set P of p_{ijk}. Then the likelihood equation for p is given by:

$$\frac{\partial G}{\partial p}=\frac{n_p}{p}+\lambda_1=0,$$

where n_p is the cell count n_{ijk} corresponding to element p.

$$\therefore p=-\frac{n_p}{\lambda_1} \qquad (5.3.3).$$

(ii) $Q=\{p_{101},p_{010}\}$:

Let q be a general element of the set. Then the likelihood equation is given by

$$\frac{\partial G}{\partial q}=\frac{n_q}{q}+\lambda_1+\lambda_2=0.$$

$$\therefore q=-\frac{n_q}{\lambda_1+\lambda_2} \qquad (5.3.4).$$

(iii) $R=\{p_{110},p_{001}\}$:

Similarly, let r be a general element of the set. The likelihood equation is given by

$$\frac{\partial G}{\partial r}=\frac{n_r}{r}+\lambda_1-\lambda_2=0$$

$$\therefore r=-\frac{n_r}{\lambda_1-\lambda_2} \qquad (5.3.5).$$

The conditions (a) and (b) from above are now applied to the estimates of p_{ijk} for each i, j and k.

Condition (a) gives:

$$-\left(\frac{n_{111}}{\lambda_1}+\frac{n_{100}}{\lambda_1}+\frac{n_{011}}{\lambda_1}+\frac{n_{000}}{\lambda_1}+\frac{n_{101}}{(\lambda_1+\lambda_2)}+\frac{n_{110}}{(\lambda_1+\lambda_2)}+\frac{n_{001}}{(\lambda_1-\lambda_2)}+\frac{n_{010}}{(\lambda_1-\lambda_2)}\right)=1$$

Put $k_1=n_{111}+n_{100}+n_{011}+n_{000}$,

$k_2=n_{101}+n_{010}$,

$k_3=n_{110}+n_{001}$.

Then condition (a) reduces to:

$$-k_1(\lambda_1^2-\lambda_2^2)-k_2\lambda_1(\lambda_1-\lambda_2)-k_3\lambda_1(\lambda_1+\lambda_2)=\lambda_1(\lambda_1^2-\lambda_2^2)$$

(5.3.6)

Condition (b) gives:

$$\frac{n_{101}}{\lambda_1+\lambda_2}+\frac{n_{010}}{\lambda_1+\lambda_2}=\frac{n_{110}}{\lambda_1-\lambda_2}+\frac{n_{001}}{\lambda_1-\lambda_2}$$

$$\therefore n_{101}(\lambda_1-\lambda_2)+n_{010}(\lambda_1-\lambda_2)=n_{110}(\lambda_1+\lambda_2)+n_{001}(\lambda_1+\lambda_2)$$

$$\therefore \lambda_2=k_4\lambda_1 \qquad (5.3.7),$$

where $k_4=\dfrac{(n_{101}+n_{010}-n_{110}-n_{001})}{(n_{101}+n_{010}+n_{110}+n_{001})}$.

Combining (5.3.6) and (5.3.7) gives:

$$\lambda_1^{\,3}(1-k_4^{\,2})+\lambda_1^{\,2}[k_1(1-k_4^{\,2})+k_2(1-k_4)+k_3(1+k_4)]=0$$

$$\therefore \hat{\lambda}_1=-\frac{[k_1(1-k_4^{\,2})+k_2(1-k_4)+k_3(1+k_4)]}{(1-k_4^{\,2})},$$

$$=-\left(k_1+\frac{k_2}{(1+k_4)}+\frac{k_3}{(1-k_4)}\right) \qquad (5.3.8).$$

But we can express:

$$k_4=\frac{(n_{101}+n_{010}-n_{110}-n_{001})}{(n_{101}+n_{010}+n_{110}+n_{001})}=\frac{k_2-k_3}{k_2+k_3}.$$

giving,

$$(1+k_4)=\frac{2k_2}{(k_2+k_3)},$$

$$(1-k_4)=\frac{2k_3}{(k_2+k_3)}.$$

Thus (5.3.8) reduces to:

$$\hat{\lambda}_1=-(k_1+k_2+k_3)=-N.$$

and hence $\hat{\lambda}_2=-k_4N.$

Substituting for $\hat{\lambda}_1$ and $\hat{\lambda}_2$ into equations (5.3.3) to (5.3.5) gives the *ML* estimates for the parameters p_{ijk} from the three sets P, Q and R:

$$p=\frac{n_p}{N},$$

$$q=\frac{n_q}{(1+k_4)N}=\frac{n_q(k_2+k_3)}{2k_2N}$$

$$r=\frac{n_r}{(1-k_4)N}=\frac{n_r(k_2+k_3)}{2k_3N}.$$

where k_2, k_3 and k_4 are constants as defined above.

To test the null hypothesis that the two classifiers S_1 and S_2 have equal error rates, we use the *deviance* given by:

$$D = -2\left(L^{(2)} - L^{(1)}\right),$$

where $L^{(1)}$ and $L^{(2)}$ are the log-likelihoods under conditions (1) and (2) from above. Using the approximation to the deviance for the multinomial distribution,

$$D \sim \chi_{n-p}^{2},$$

where n is the number of cells and p is the number of independent unknown parameters. In this problem $n = 8$ and $p = 5$. Therefore, we reject the null hypothesis that the two classifiers give equal error rates if the observed value of D is greater than the upper $100\alpha\%$ point of the χ_3^2 distribution, where α is the *significance* level.

5.3.2.2 Comparison of bad rates

Similar likelihood statistics can be derived to compare the bad rates under two classifiers S_1 and S_2. In this case we ignore the cells (0,0,0) and (1,0,0) because we are only concerned with applicants who are accepted using at least one classifier. The two conditions under which we compare the log-likelihood are:

(1) Assuming no relationship between parameters.

(2) Assuming that the expected probability of classifying a bad as an accept is the same under both classifiers. Using the above notation this can be expressed as

$$p_{010} + p_{011} = p_{011} + p_{001},$$

which reduces to

$$p_{010} = p_{001} \qquad (5.3.9).$$

As before this condition represents the null hypothesis of no difference between classifiers.

Because of the reduced number of parameters, the ML estimation of the log-likelihood is easier in this case. In the unconstrained case, condition (1), the parameter estimates are

$$\hat{P}_{ijk} = \frac{n_{ijk}}{N},$$

where i,j and k range over the reduced set of possible values and N is redefined accordingly.

The ML estimates corresponding to the constrained model, condition (2), are easily estimated by substituting condition (5.3.9) into the log-likelihood and maximising subject to the usual condition that the cell probabilities sum to 1. The resulting log-likelihood is given by:

$L = n_{111}\log(p_{111}) + n_{110}\log(p_{110}) + n_{101}\log(p_{101}) + n_{011}\log(p_{011}) + (n_{010} + n_{001})\log(p_{001})$

and the constraint is $p_{111} + p_{110} + p_{101} + p_{011} + 2p_{001} = 1$ (5.3.10).

The Lagrange multiplier is give by:

$$F = L + \lambda(p_{111} + p_{110} + p_{101} + p_{011} + 2p_{001} - 1).$$

The corresponding likelihood equations are given by:

$$\frac{\partial F}{\partial p_{ijk}} = \frac{n_{ijk}}{p_{ijk}} + \lambda = 0 \text{ for } (i,j,k) \neq \{(0,1,0),(1,0,0),(0,0,0)\}$$

and

$$\frac{\partial F}{\partial p_{001}} = \frac{(n_{001} + n_{010})}{p_{001}} + 2\lambda = 0.$$

Rearranging these equations gives:

$$p_{ijk} = -\frac{n_{ijk}}{\lambda}$$

and

$$p_{001} = -\frac{(n_{001} + n_{010})}{2\lambda}.$$

Applying (5.3.10):

$$\lambda = -(n_{111} + n_{101} + n_{110} + n_{011} + n_{010} + n_{001}) = -N.$$

The resulting parameter estimates are given by:

$$p_{ijk} = \begin{cases} \dfrac{(n_{001} + n_{010})}{2N} & (i,j,k) = \{(0,0,1),(0,1,0)\} \\ \dfrac{n_{ijk}}{N} & \textit{otherwise} \end{cases}.$$

As before the test statistic is the deviance between the log-likelihoods with parameters estimated under the conditions (1) and (2) considered above. The null hypothesis of equal bad rates is tested by comparing the observed value of D with the appropriate point of the χ_3^2 distribution. This likelihood ratio test is used to compare the relative performance of two classifiers in later chapters of the thesis (see for example Section 7.2.3).

5.4 Conclusions

For commercial reasons the criterion adopted in this thesis is the minimisation of the bad rate amongst the accepts, given a fixed acceptance rate. This criterion is unusual outside the credit scoring field and has implications for the levels of performance that can be achieved. In Section 5.2.2 we discussed the bounds on performance that result from fixing the acceptance rate.

In this chapter we presented a general review of other approaches to measuring the performance of a classifier. In particular, we distinguished between absolute and relative performance. Absolute performance was further sub-divided into measures of discriminability and reliability. Discriminability measures assess how successful a classifier is in allocating applicants to their true class, whereas reliability measures assess the accuracy of the estimates of $P(g|\mathbf{x})$.

The bad rate is a particular type of discriminability measure closely allied to the error rate. It is less sensitive than continuous measures of performance, described in Section 5.2.3.2, but may be more appropriate when the objective of the system is to minimise bad debt. In Section 5.2.3.3, we discussed another criterion for choosing between different discriminability measures, the idea of properness. Using this criterion the bad rate would be rejected in favour of continuous measures such as the Brier score, C_2.

In Section 5.2.4 we discussed an approach to constructing reliability measures from individual discriminability measures described by Hilden et al. (1978). Although the approach taken is appealing, we identified several weaknesses. Moreover, reliability is generally of less interest to a credit grantor than discriminability.

In Section 5.3 we proposed two tests for assessing the relative performance of two classifiers. Both tests can be used to compare bad rates or other measures based upon the counts of misclassification (see Section 5.2.3.1). The two tests are complementary and can be used together to decide whether one classifier is better than other. In particular the two tests address slightly different questions.

- The first test compares the bad rate amongst applicants exclusively accepted under one of the two classifiers. By excluding applicants accepted under both classifiers, any small but real difference in the overall bad rates is more likely to be identified as significant.

- The second test uses a likelihood based approach to compare the overall bad rate from two classifiers. Because all accepted applicants are included in the test, it is more appealing from a conceptual point of view. This also makes it more conservative in judging a difference in bad rate to be significant. As a result there is less danger of mistakenly identifying a difference as significant when it can be attributed to random fluctuation.

Chapter 6

Reject Inference

6.1 Introduction

In Chapter 2 we described how applicants for credit are assessed using a credit scoring model leading to a decision to accept or reject them. Because the rejected applicants are not granted credit, their true creditworthiness cannot be determined. *Reject inference* is the process of trying to infer the true credit status of the rejects, using their characteristic vectors.

Two important reasons for needing reject inference are described below:

(1) Having developed a scorecard at some previous time, a credit grantor may be interested in whether a better scorecard can be constructed using the current information available on the accepts and rejects. One reason for wanting to do this is degradation of scorecard performance due to population drift: the tendency for populations to evolve over time. This problem is mentioned in a medical context in Hand and Henley (1994) and ways of detecting population drift are suggested. In the credit environment the state of the economy is probably the most important factor, but undoubtedly changes in demographic characteristics, fashions, and attitudes will play a part.

One problem that arises when building new scorecards is that, as mentioned above, the true good/bad status for the rejects is unknown. If a sample consisting only of accepted applicants is used to construct the new scorecard then bias may be introduced. Traditionally reject inference has been used to try and reduce this bias by inferring the true status of the rejects and combining this with the known status of the accepts to produce a new scoring instrument. However, much of the published work on reject inference seems to have been based upon a poor understanding of what it is possible to achieve using the characteristic vectors for the rejects. This is an important methodological issue which needs to be addressed before constructing a credit scoring system.

(2) Another motivation for using reject inference is to obtain an accurate estimate of the proportion of potential goods (from the full applicant population) being rejected by the existing scorecard. Other unknown aspects of the structure of the reject population may provide further reasons for needing reject inference (for example, if we wish to know the shape of the $P(g|\mathbf{x})$ curve in the reject region).

In this chapter we will concentrate on the first aim described: using reject inference to update scorecards constructed on the accepts sample in order to reduce the sample selection bias. We will precede this by a consideration in Section 6.2 of the nature and size of the bias. We had access to a *validation* sample which included the true creditworthiness of the rejects and this allowed us to compare models built on the full and accept samples to confirm the existence of bias. The validation sample allowed us to assess all the work considered in this chapter.

Reject inference is a technique that is widely used by scorecard developers, although it has not been given extensive treatment in the literature. In Section 6.3 we will discuss the methods that have been proposed.

Having considered the relevant literature we will proceed to describe the various approaches to reject inference that we have taken. In Section 6.4 we consider a first approach, extrapolation from a model built on the accepts sample. We will see how this term can be used to describe a variety of methods and examine factors that may affect their success.

In Section 6.5 we consider whether the characteristic vectors for the rejects can be used to develop improved scoring rules. We present an analysis of likelihood functions to show that the rejects' vectors do not contain information about the parameters of the observed data likelihood. Therefore, one is not able to produce a better scorecard than one obtained using the accepts sample, except by chance or if extra information is incorporated in some way. This has important implications for developers of credit scoring systems using reject inference.

The remainder of the chapter will focus on ways of incorporating additional assumptions or supplementary information into methods of reject inference. Section 6.5.4 considers how assuming distributions for the goods and bads separately allows use to be made of the rejects' characteristic vectors. This apparent gain in information needs to be balanced with the reasonableness of the assumptions made. In Section 6.6 we look at the use of supplementary information in the form of a subsample of cases from the reject region with true status known. We call the sample used, for which the true status is known for accepts and rejects, a *calibration* sample. Several methods of reject inference which use the calibration sample are proposed and evaluated.

In Section 6.6.5 we present a novel approach which uses *foresight* data. This is data that has been collected on applicants since they first applied for credit (for instance, an applicant may receive a county court judgement from another debt after being rejected by the credit grantor who wishes to perform reject inference). This information obviously cannot be incorporated into the scorecard, because it is not available at the time an applicant applies for credit, but it will have a strong link with the true status of an individual. Two methods of using this information to provide more accurate estimates of the good/bad probabilities for the rejects are considered.

In the final section of this chapter we present an empirical comparison of some of the reject inference methods considered and draw conclusions about the best approach to reject inference.

6.2 Is reject inference necessary?

Eisenbeis (1978) reports that using a model based solely on the truncated population of accepted applicants can frequently generate misleading results. Other authors, including Hsia (1978), Reichert et al. (1983) and Joanes (1993/4) have highlighted the possible bias that can result from using a sample of accepted applicants to build a scorecard with which to assess the full applicant population. However, little attention in the literature has been directed to evaluating the factors which affect whether reject inference is necessary and identifying the causes of the bias One reason for this is that the

credit scoring developer does not usually have access to information about $P(g|\mathbf{x})$ in the reject region. In this section we explore these issues with the aid of a *validation* sample from a previous time period, which includes the true creditworthiness of the rejects.

The term *bias* is used to refer to the difference in performance (in terms of bad rate) between the full and accept sample scorecards. A scorecard is said to be biased if it differs in bad rate from the full sample scorecard. We address this bias indirectly by considering the bias of the class membership probabilities for the accepts. The accepts sample is said to be biased at a point in the characteristic space if the proportion good at this point is different for the full and accept samples. The validation sample is used to explore explore some explanations for this bias.

In Section 6.2.1 we present comparison studies of the relative performance of classifiers constructed using the full and accept portions of the validation sample. Inferences are made about some of the factors influencing the relative performance of a scorecard built on the accepts sample. We shall see that there are situations in which reject inference is unnecessary. However, this is a very dangerous assumption to make if no validation sample is available to confirm it. In the majority of cases the accepts scorecard will be biased and so some form of reject inference is desirable.

In Section 6.2.2 we consider a standard approach (in the credit industry) to explaining the bias of a scorecard built on the accepts. This involves the discussion of characteristics, called *derogatory characteristics*, which may not be given sufficient weighting in models built on the accepts.

In Section 6.2.3 we describe several approaches to explaining the bias of linear regression classifiers when the data is represented by weights of evidence. A simple one characteristic example is used to assess their contribution. In particular, we emphasise the importance of the set of characteristics used to design the new classifier. We show that if this set does not include all the characteristics used to make the original accept/reject decision, then the accepts classifier will be biased. The validity of the assumption of a linear relationship between $P(g|\mathbf{x})$ and weights of evidence transformation is also considered.

Additive and multiplicative models for the bias of the accept sample are considered for the one dimensional case in Section 6.2.4.

Further discussion of these issues is presented in Section 6.4, where we consider extrapolation from the accepts as an approach to reject inference. There we discuss the effect of different approaches to classifier design on the performance of extrapolation methods.

6.2.1 Comparison of accept and full sample classifiers

A first analysis was carried out using the validation sample described in Table 2.1. A maximum of twenty characteristics were available for both the full and accept samples. To give an idea of how the supposed bias effect changes as the proportion of accepts in the original sample changes, the analysis was carried out with two different accepts samples. The first accepts sample came from accepting the "best" 70% of the original full sample (where "best" is determined by scoring up the sample under a previous scorecard) and the second sample came from accepting the best 50% of the full sample.

Scorecards were constructed on the full and accept validation samples using linear regression with the data in weights of evidence form. Characteristics were selected using a combination of the methods described in Section 2.3. The bad rates among those scoring above the threshold in the independent test set are shown for different acceptance rates in Table 6.1. The overall bad rate in the test set is 31.34%.

	Acceptance rate		
Design Sample	30% accepts	50% accepts	70% accepts
FULL	5.98	12.68	19.69
ACCEPTS SAMPLE (70%)	6.09	12.76	20.23
ACCEPTS SAMPLE (50%)	5.89	13.30	21.08

Table 6.1: Bad rates at different thresholds for scorecards built using full and accepts samples.

The bad rates for the different scorecards in Table 6.1 can be compared using the significance tests described in Chapter 5. The results are significantly

different at the 70% acceptance threshold with a significance level of 10%. Because a small reduction in bad debt can result in a large saving for the credit grantor, it can be argued that we should employ a significance level that allows any real difference between two scorecards to be picked up (a high power). The significance level of 10% was fixed in order to facilitate this.

Several additional points are suggested by the preceding results. First, when there is a real difference between the performance of the full and accept scorecards, the difference in terms of the bad rate appears to be quite small. Thus, there is a small target range for the performance of reject inference methods (because we would not realistically expect to be able to do better than the full scorecard). This makes the problem hard to solve, but it is still one of importance because of the savings and increased profits that could be made if even a small reduction in bad rate was achieved.

A related point is that the difference between the accept and full scorecards depends on the proportion of the full sample used to make the accepts sample. The higher the proportion of accepts used to build a scorecard, the less the bias in the bad rate. This should be taken into account when a credit scoring system developer is considering how to proceed with reject inference. If only 30% of the applicant population are accepted then some form of reject inference is likely to be essential to reduce the bias effect.

Thirdly, the accept/reject mechanism affects how representative the accept sample is of the full sample, and can thus affect whether there is any real difference between the full and accept scorecards. In the example presented above the classifier that was used to split the sample into accepts and rejects was provided by the mail order company and achieves good discrimination between goods and bads. Therefore, the accepts sample contains a particularly low proportion of bads and is likely to weight incorrectly some of the attributes of a bad credit risk. This will increase the possible difference between the full and accept scorecards.

The conclusion that we can draw from Table 6.1 is that, in some circumstances, there is a significant difference between scorecards built on the full and accept samples. In these cases, the rejects with their true classes (which are unknown in practice) do add vital information for discriminating between good and bad

applicants in the full population. This necessitates the search for appropriate methods of reject inference.

A second analysis was carried out to confirm the effect of the accept/reject mechanism and the truncation points on the performance of accept sample models. These factors were controlled by:

- Reducing the number of characteristics available for constructing the original classifier. This was to decrease the discriminatory power of the classifier used to make the accept/reject decision. (Twelve and fifteen characteristics were used.)

- Fixing the truncation point so as to accept 70% of the full sample.

Fixing these two factors at the above levels should reduce the difference between new classifiers built on the full and accept samples. Table 6.2 shows the bad rates obtained at a threshold giving a 70% acceptance rate for an independent test set.

	12 char's	15 char's
Full	20.69	20.63
Accepts	20.77	20.77

Table 6.2: Bad rates for the full and accept scorecards built using linear regression.

There is not a significant difference between the bad rates for the full and accept samples (using the significance tests described in Section 5.3), suggesting that in this case it is not necessary to perform reject inference. It is not possible to improve on the scorecard built using the accept sample. The results confirm our prediction and show that reject inference becomes more important as the original accept/reject classification becomes better.

We have seen in the second example that the accepts sample may not always be sufficiently biased to warrant reject inference. Even if this is not the case, then it may still be more economical for the scorecard developer to use the accepts for scorecard construction rather than giving credit to a sample of rejected

applicants in order to build a representative sample. As we shall see later, reject inference has an associated cost in terms of the need for additional, possibly expensive, information.

In practice a scorecard developer will probably not have access to a validation sample with the true creditworthiness known for the rejects. Thus, they will not be able to carry out the sort of analysis described above in order to determine whether reject inference is necessary or not. This may explain the reason why it is often assumed that reject inference is necessary together with claims about the effectiveness of particular methods of reject inference (see Section 6.3).

From the above it is apparent that bias can sometimes occur when constructing scorecards using only the accept sample. We will now consider ways of trying to explain the nature of the bias and examine whether this gives any indications as to sensible ways to proceed with reject inference.

6.2.2 Derogatory characteristics

Some developers of credit scoring models explain the need for reject inference using the idea of a *derogatory characteristic.* An informal definition of a derogatory characteristic is one which has attributes which do not distinguish between goods and bads in the accepts sample, but give good discrimination between goods and bads in the full population. The argument goes that, for this reason, the derogatory characteristic will not get sufficient weighting in a scorecard designed on the accepts sample. (Joanes (1993/4) uses a version of this argument, without using the term derogatory characteristic, to justify the need for reject inference.)

A common example of a derogatory characteristic (see Table 6.3 for details) is "Weeks since last county court judgement". There are two reasons why this satisfies the above definition: first, whether an applicant has had a previous county court judgement or not is highly correlated with creditworthiness and so this characteristic discriminates well between good and bad applicants in the full sample. Secondly, because of the first factor, this characteristic is likely to be used in the original accept/reject classifier, thus screening out a high proportion

of applicants with CCJs. This means that the characteristic will lose its discriminatory power when the accepts sample is considered alone.

There are two basic properties of derogatory characteristics which, it is argued by credit scoring practitioners, introduce bias into scorecards built on the accepts (we examine these properties in more depth in Section 6.2.3):

- There are too few applicants from the accept sample in attributes of the derogatory characteristic with low weights of evidence (e.g. few accepts are likely to have county court judgements). As a consequence random variation may cause there to be an artificially low (or high) number of bads from the accepts in these "bad" attributes. Therefore, if we build a scorecard using the accepts then the resulting values assigned to the bad attributes may be unreliable.

- Attributes with low weights of evidence in the full sample may contain an overly high number of goods in the accepts sample due to bias in the sample. There are likely to be some good applicants who have one or more derogatory attributes (e.g. a previous default due to extenuating financial circumstances) and this may give an under-representative impression of the proportion of bads in the full sample with this attribute. The result of this property is that when using the accept sample to build new scorecards, simple linear interpolation overestimates the proportion of goods for low values of the total score.

These properties are used to justify why a scorecard built on the accepts sample may be biased. In order to reduce this bias it is necessary to adjust the proportions and predicted numbers of goods and bads in attributes with low weights of evidence. Ad hoc methods of reject inference which use this approach could explain the claims of using the rejects to improve a scorecard. In Section 6.5.3 we show that there is not an objective way of using the characteristic vectors for the rejects to reduce bias.

6.2.3 Explanations for bias

In this section we consider explanations for the bias of linear regression classifiers built on the accept sample. It is assumed that the data is represented

by weights of evidence. We attempt to explain the underlying mechanisms which introduce the bias and prescribe ways of reducing it. (For more general discussion of extrapolation methods and the causes of bias see Section 6.4.)

We distinguish between two aspects of the original accept/reject decision which can contribute to the bias. We then investigate their relative importance with the aid of some simple examples:

(1) *The sampling fraction used to select the accepts sample.*
If a characteristic has a nonlinear relationship with $P(g|\mathbf{x})$, then by taking a sampling fraction which varies arbitrarily across the attribute values, a linear regression model built on the accepts sample will be biased. (We call this *factor A*) We examine the relationship between weights of evidence and proportion good for different characteristics to identify whether the relationship is non-linear.

A further cause of bias (*factor B*) is the random variation in the sample proportion good as the sampling fraction varies. This is equivalent to the first property of derogatory characteristics outlined above.

(2) *The set of characteristics used to make the accept/reject decision.*
If the accept/reject decision uses extra information, not available for building new classifiers, then the proportion of goods among the accepts will typically not equal the proportion of goods amongst the rejects at any particular point in the characteristic space. This could lead to bias in a model built on the accepts (*factor C*).

To simplify the following analysis, we discuss the problem of building a classifier using one characteristic, "Weeks since last county court judgement (C.C.J)". Table 6.3 gives the attribute values, the proportions of goods and bads in the full and accept portions of our validation sample and the corresponding weights of evidence. The accepts sample comes from applying a previous (and now unavailable) classifier to the full sample, using a range of characteristics.

	Attribute	Goods	Bads	$P(g)$	Weights
Full	1 (1-26 wks)	242	825	0.227	-1.127
	2 (27-52)	253	578	0.304	-0.726
	3 (53-104)	340	742	0.314	-0.681
	4 (105-208)	633	1043	0.378	-0.4
	5 (209-312)	509	686	0.426	-0.199
	6 (No CCJ)	5495	4382	0.556	0.326
Accepts	1	83	103	0.446	-0.898
	2	68	77	0.469	-0.806
	3	123	127	0.492	-0.714
	4	301	244	0.552	-0.472
	5	301	176	0.631	-0.145
	6	4743	2114	0.692	0.126

Table 6.3: Description of the characteristic "Weeks since last CCJ".

The first stage is to consider plots of proportion good against weight of evidence for the full and accept samples. Figures 6.1 to 6.3 show such plots using weights of evidence calculated from the full sample, the accept sample and the appropriate sample (full or accepts) respectively. 95% confidence intervals are added to all the plots to give an impression of how reliable the proportions are. For an attribute that has $P(g|\mathrm{x})=\pi$ and a total of m agents, an approximate 95% confidence interval is given by $(\hat{\pi}-2\sqrt{\hat{\pi}(1-\hat{\pi})/m}$, $\hat{\pi}+2\sqrt{\hat{\pi}(1-\hat{\pi})/m})$. This assumes the normal approximation to the binomial distribution for large values of m. Because the full sample is considerably larger than the accepts sample, the corresponding probability estimates are more accurate and so the confidence intervals are smaller.

The above plots show that, given a particular sample, the proportion good has an approximately linear relationship with weight of evidence when the weights of evidence are calculated from that sample. In fact, the relationship can be worked out explicitly. If an attribute has proportion good p and the total number of goods and bads in the sample are equal, then the weight of evidence for that attribute can be expressed as $\log\frac{p}{(1-p)}$. The resulting curve of proportion good against weight of evidence is approximately linear for

Fig 6.1: Proportion good using full sample weights of evidence

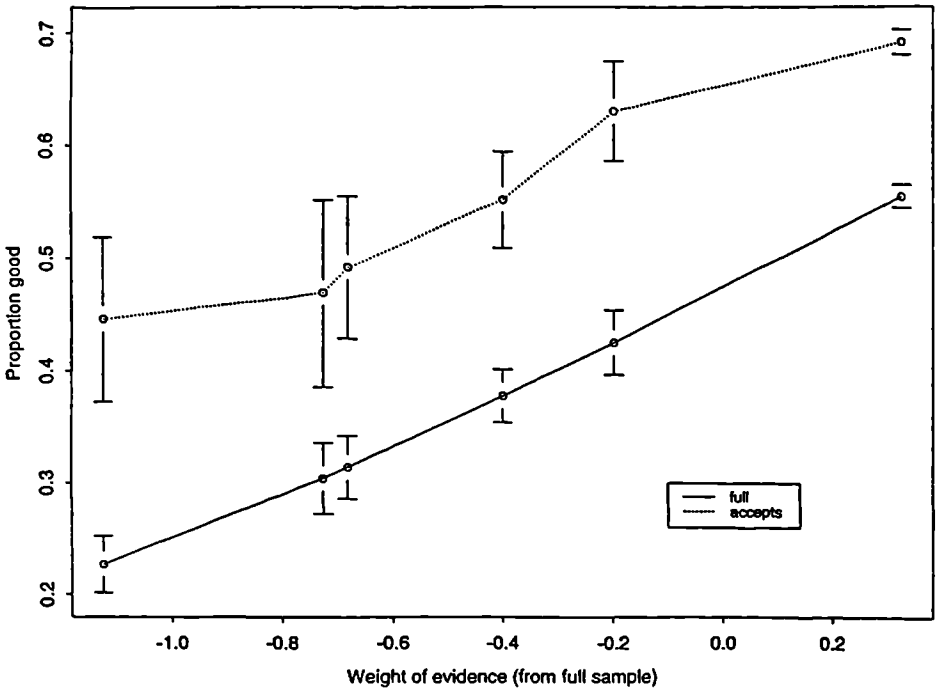


Fig 6.2: Proportion good using accepts sample weights of evidence

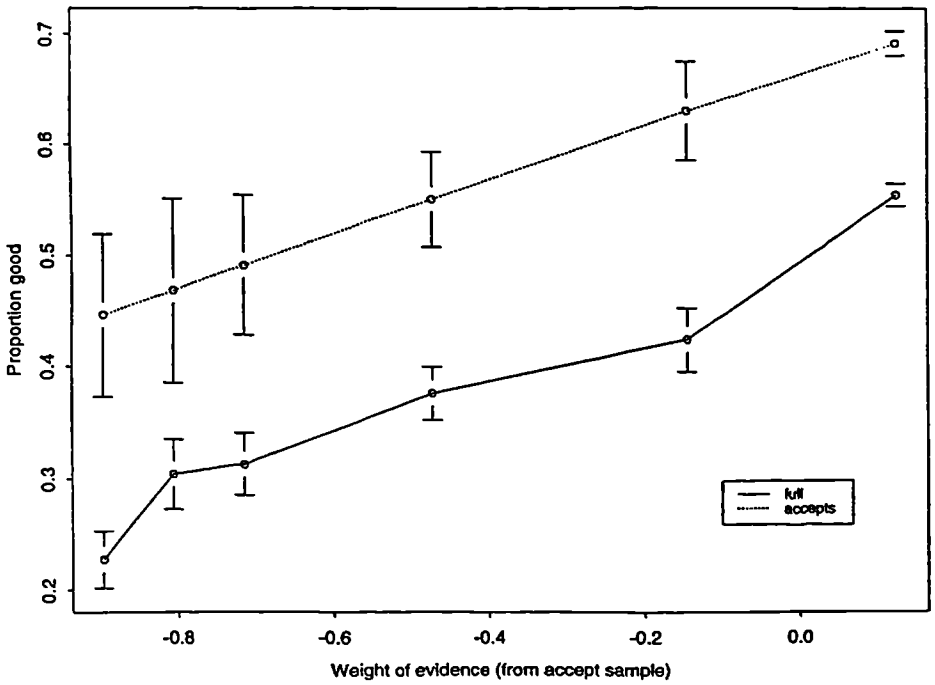


Fig 6.3: Proportion good using appropriate sample weights of evidence

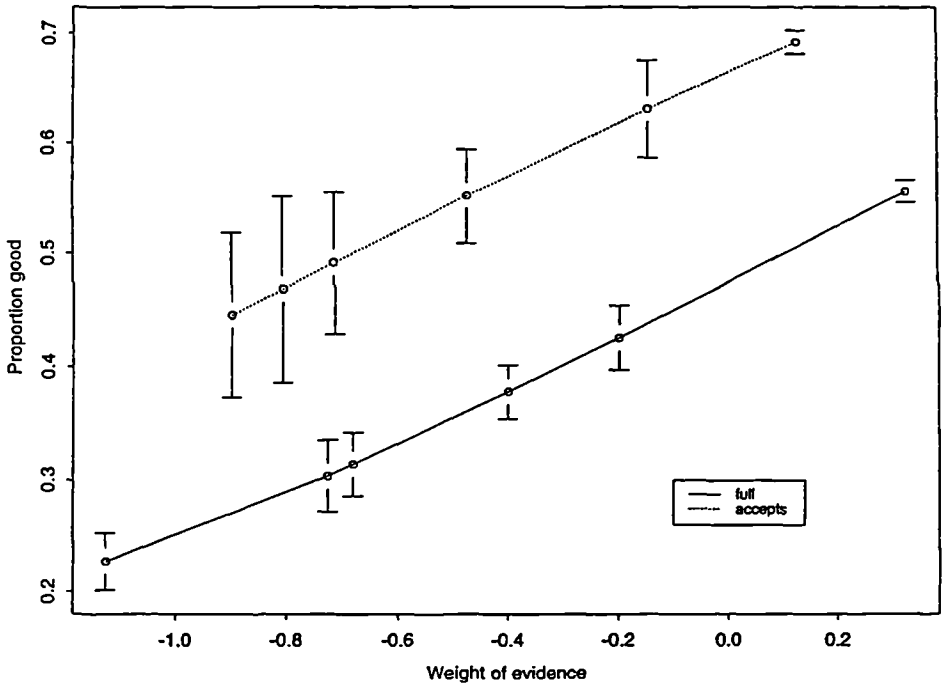


Fig 6.4: Proportion good against full sample weights of evidence for postcode characteristic

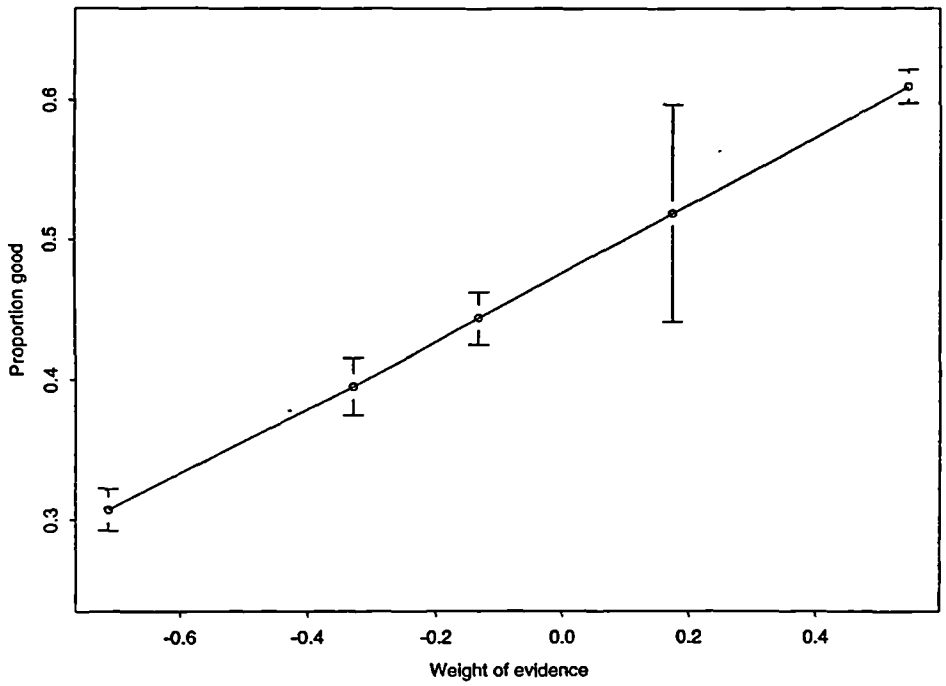


Fig 6.5: Proportion good against full sample weights of evidence for decision tree characteristic

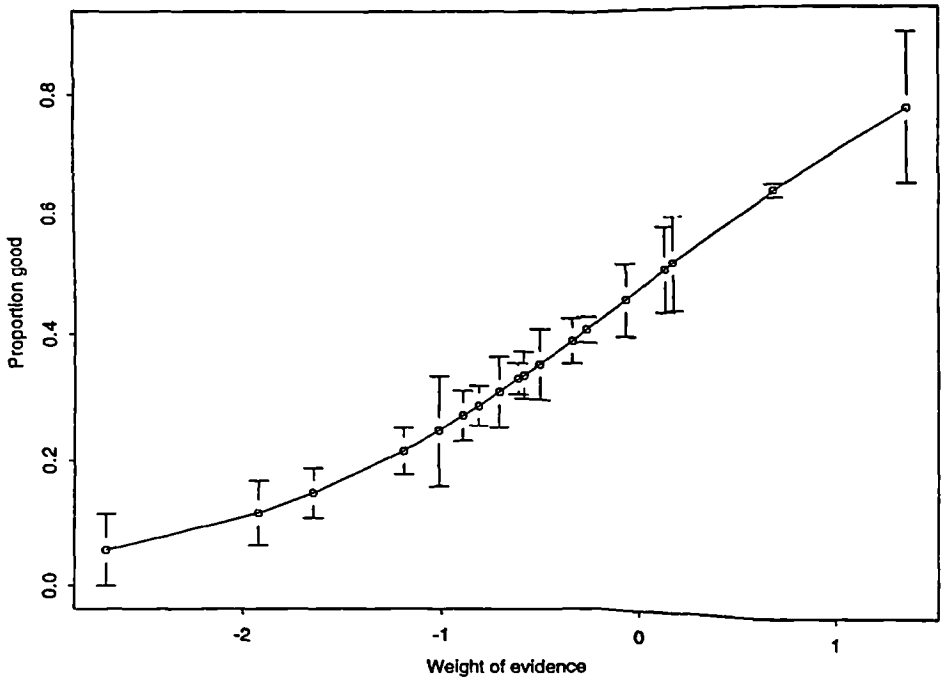
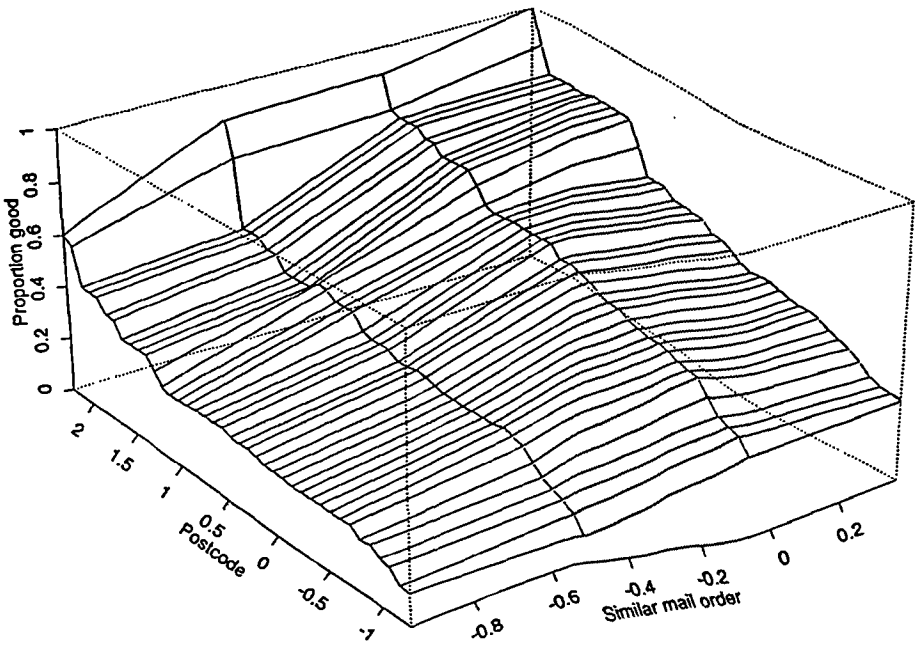


Fig 6.6: An example where the surface of P(g) is not linear



$0.2 \leq p \leq 0.8$. (However, when the weights of evidence are calculated from a different sample the relationship may not be approximately linear.)

It was found that for the majority of characteristics the relationship between weight of evidence and proportion good was approximately linear. However, for some characteristics the relationship was noticeably non-linear. This was dependent on the range of values of the proportion good, p, as described above. For example, Figures 6.4 and 6.5 show one-dimensional plots for a decision tree characteristic (non-linear) and a characteristic combining postcode with time at address (linear). In both cases the full sample curve is shown with full sample weights of evidence.

Figure 6.6 shows a two-dimensional example with proportion good plotted against weights of evidence for a postcode characteristic and a characteristic describing previous mail order history. The values of proportion good were smoothed using a kernel function to reduce sample variation. The resulting surface is non-linear. In fact, departures from linearity are likely to increase as the number of dimensions increases. This is because the weights of evidence transformation is one-dimensional (by transforming the attribute values according to the numbers of goods and bads from the marginal distributions). As the dimensionality increases the interactions between characteristics distort the approximately linear relationship with proportion good.

We conclude that in high dimensions *factor A* will be a significant cause of bias in classifiers built on the accepts. However, in low dimensions (one or two characteristics) the weights of evidence provide a suitable means of transforming the characteristic values onto an approximately linear scale. In this situation *factor A* is not an important component of the bias.

In order to understand the contribution of *factors B* and *C* to the bias, we return to the one-dimensional problem of building a classifier for the characteristic "Weeks since last CCJ". Table 6.3 showed that the proportion of the full sample in the accept sample varies according to attribute. To make this clearer, the second column of Table 6.4 shows the sampling fraction, S_1, for the accepts (it is given by the proportion of the full sample accepted by the original classifier within each attribute).

Attribute	Sampling fraction S_1	S_2
1	0.174	0
2	0.174	0
3	0.231	0
4	0.325	0
5	0.399	0
6	0.694	0.857

Table 6.4: The sampling fraction for the characteristic "Weeks since last CCJ".

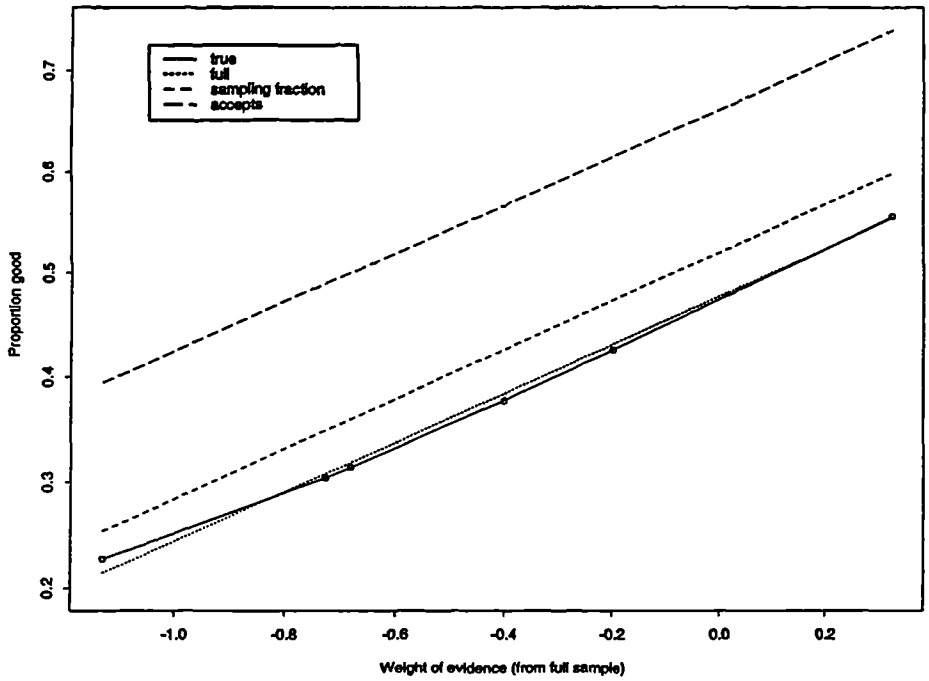
The third column shows the sampling fraction, S_2, which would be used to accept the same proportion (53.8%) of the full sample if this was the only characteristic available.

Figures 6.1 to 6.3 indicate that for this characteristic there is an approximately linear relationship between weight of evidence and proportion good. Therefore, any bias caused by varying the sampling fraction can be attributed to *factor B*. Furthermore, if a classifier is constructed using the portion of the full sample obtained using sampling fraction S_1, the resulting bias will be independent of *factor C*. This is because no extra information/characteristic is used to make the accept/reject decision. This gives us a way of assessing the contribution of *factor B* to the bias.

The contribution of *factor C* can be identified by building a classifier using the same sampling fraction S_1 from the full sample, but including extra information in the decision. The accepts sample that we have available satisfies these conditions. The resulting contribution of *factor C* comes from the difference in bias between the two classifiers described.

The above approach was utilised by constructing simple linear regression models using three different sampling methods: (1) using the full sample, (2) using the sub-sample obtained from the full sample with fraction S_1 and (3) using the accepts sample. Figure 6.7 shows the regression lines for the three methods described. The full sample mean values are added for each of the

Fig 6.7: Regression lines built using full and accept samples



attributes using the full sample weights of evidence to represent the "true" values of proportion good.

The above plot indicates that there is a difference in level between the three regression lines, but that they all roughly parallel. This indicates that the bias is approximately constant across attribute values. (We present further investigation into how the bias changes with attribute value in the next section.) The bias from *factor B* is represented by the difference between the full and sampling fraction regression lines. The bias from *factor C* is represented by the difference between the sampling fraction and accept sample lines. Both these factors are contributing to bias of the regression model.

Performance of the methods was assessed more formally using the bias measure:

$$B = \frac{1}{n}\sum_{i=1}^{n}(y_i - r_i)^2$$

where y_i is the true class of the ith applicant and r_i is the regression estimate for the ith applicant. The resulting values of the bias are shown in Table 6.5.

Method	Bias
1	0.0000217
2	0.00189
3	0.0339

Table 6.5: Bias of regression lines constructed from "Weeks since last CCJ".

The table confirms that both *factors B* and *C* contribute to the bias of classifiers built on the accepts sample. In this example *factor C* has a much bigger effect.

To conclude, we have identified three causes of bias in accept sample models using linear regression. In particular we have found:

(1) The assumption of a linear relationship between weight of evidence and proportion good ($P(g|\mathbf{x})$) may not be appropriate in high dimensions. Bias is introduced by classifiers which vary the sampling fraction across the characteristic space (*factor A*).

(2) The weights of evidence transformation may provide a linear relationship between attribute value and proportion good in low dimensions. Bias can still result from random variation in the sample (*factor B*).

(3) An important cause of bias is the set of characteristics used to build the new classifier. If the accept/reject decision is made using a set of characteristics X and a new scorecard is to be constructed using a set of characteristics, Y, then an accepts scorecard will generally be biased if $Y \subset X$ (see also Hand and Henley, 1993/4). The same argument applies if the new scorecard uses a set of characteristics, Z, where Z does not include all of X. Therefore, in order to reduce the bias in an accepts classifier it is necessary to include all the characteristics used to make the original accept/reject decision.

(4) We have seen that there may be a difference in level between the good/bad proportions for the full and accept samples. (This corresponds to the regression lines being parallel.) However, this form of bias may not, in itself, lead to inferior classification performance. This is because imposing suitable thresholds on parallel probability surfaces will lead to the same applicants being accepted under both classifiers. On the other hand, if the bias changes with attribute value (when the regression lines are not parallel or monotonically related) then imposing a threshold may lead to a change in performance.

(5) We can use the results of this section to explain the two properties of derogatory characteristics described in Section 6.2.2. The first property of derogatory characteristics is equivalent to the bias due to random sampling which results from taking an arbitrary sampling fraction across the characteristic space (*factor B*). The second property of derogatory characteristics results from using extra information, not available for building new classifiers, to make the original acceptance decision (*factor C*). We note that any characteristic which has attributes which differentiate between good and bad applicants in the full sample can have these properties.

In Section 6.2.4 we give further attention to the nature of the bias. Formal tests are used in the one-dimensional case to assess whether the accept sample bias changes with attribute value. We consider both additive and multiplicative models.

6.2.4 Additive and multiplicative models of bias

We present models for bias in the accept sample when the original acceptance decision uses information not available for constructing new classifiers. For simplicity we continue to consider the problem of constructing a classifier with one characteristic, "Weeks since last CCJ" (see Table 6.3 for details).

The bias of the accept sample can be split into two types:

(1) A constant bias across attribute values. This form of bias may not lead to inferior classification performance. It is represented by the difference in level between the full and accept curves of proportion good in Figures 6.1 to 6.3.

(2) A variable bias across attribute values. This form of bias is more important than the first because it is more likely to lead to a change in ranking by $P(g|\mathbf{x})$ and, thus, a change in classification performance. It is represented by a difference in shape between the full and accept curves of proportion good. It is not clear from Figures 6.1 to 6.3 whether there is a difference in shape between the two curves for the accept sample considered.

We focus on testing for the second type of bias because it has a more significant influence on classification performance. Two ways in which the bias can vary with attribute value are discussed: additive and multiplicative relationships. Table 6.6 shows the proportion good for the full and accept samples together with the ratio of the accept to full probabilities and the difference between probabilities across attributes. Variation in the ratio corresponds to multiplicative bias and variation in the difference corresponds to additive bias.

Attribute	Full sample proportion good	Accept sample proportion good	Accept/ Full ratio	Accept - Full difference
1	0.227	0.446	1.964	0.219
2	0.304	0.469	1.542	0.165
3	0.314	0.492	1.567	0.178
4	0.378	0.552	1.460	0.174
5	0.426	0.631	1.481	0.205
6	0.556	0.692	1.245	0.136

Table 6.6: Ratios and differences of the proportion good for the accept and full samples on the characteristic "Weeks since last CCJ".

Column 3 shows clearly that the ratio of the two proportions decreases as attribute value increases. This indicates that a multiplicative model would be an appropriate way to describe the relationship between attribute value and the level of bias. Column 4 shows a fluctuating difference between the two proportions of goods. This indicates that an additive model is not an appropriate way of describing the data.

In order to formalise the previous discussion, we introduce statistical tests to evaluate the nature of the relationship between attribute value and proportion good. In order to ensure independence of the observations it is necessary to compare the accept sample with the reject sample, rather than the full sample.

(1) The additive test involves comparing the relative fit of the models given by:

$$y_{ij} = a + b_i + c_j$$
$$y_{ij} = a + b_i + c_j + d_{ij}$$

where i represents attribute value and j is 0 for the accept sample and 1 for the reject sample. The response y represents the proportion of goods in the sample.

The parameters a, b, c and d are fitted by maximum likelihood estimation and the models compared using analysis of variance to see if the second equation adds significant information about the response y through the term d_{ij} (the

interaction between attribute value and sample). The term c_j is included in both models to allow a difference in level between the two curves of proportion good. This removes the influence of the first type of bias identified above on the significance test.

(2) The multiplicative test involves comparison of the models given by:

$$\log(y_{ij}) = a + b_i + c_j$$

$$\log(y_{ij}) = a + b_i + c_j + d_{ij}$$

where the parameters are defined above. The testing procedure is similar to that for the additive model.

These tests give a formal method of assessing whether the accepts sample exhibits bias for individual characteristics. The ANOVA tables for the characteristic "Weeks since last CCJ" are shown in Tables 6.7 and 6.8.

	Df	Sum of Sq	Mean Sq	F value	P(F)
Attribute	1	0.1059	0.1059	129.4	0.0000032
Sample	1	0.0967	0.0967	118.1	0.0000045
Interaction	1	0.0006	0.0006	0.780	0.4028102
Residuals	8	0.0065	0.0008		

Table 6.7: ANOVA table for additive model of bias for "Weeks since last CCJ".

	Df	Sum of Sq	Mean Sq	F value	P(F)
Attribute	1	0.5630	0.5630	205.3	0.0000006
Sample	1	0.5412	0.5412	197.3	0.0000006
Interaction	1	0.0438	0.0438	16.0	0.0039799
Residuals	8	0.0219	0.0027		

Table 6.8: ANOVA table for multiplicative model of bias for "Weeks since last CCJ".

The tables show that for both models the individual effects for the sample and attribute value are highly significant. This confirms that proportion good varies with attribute and that there is a difference in level between the full and accept curves for proportion good. However, as expected, the significance result for the interaction effect differs between the two models. The interaction

effect is significant for the multiplicative model and non-significant for the additive model. We conclude that when the acceptance decision is taken on the basis of extra information, the accept sample is biased and the bias has a multiplicative relationship with attribute value.

The tests considered above allow us to consider the nature of the bias of the accepts sample for individual characteristics. Further work is neeed to assess the bias in the multivariate case.

6.3 Survey of reject inference methods proposed in the literature

Reject inference is a process that is used by almost all developers of credit scoring systems, but one about which there is little published work in journals on statistics or financial modelling. There does not appear to be an accepted, standard method of using the characteristic vectors for the rejects; rather each firm or consultant employs a subtly different technique that for commercial reasons they wish to keep secret. However the argument of this chapter is that there is no systematic way of doing reject inference unless extra information is incorporated in some way. (See also Hand and Henley (1993/4) and Henley and Hand (1995))

We now consider methods of reject inference that have been publicised.

Method 1

The first and most important of these, because of its longevity, is the *Augmentation* method that was first developed by the Fair Isaac company. It is described by Hsia (1978), who reports that it is widely used by developers of scoring systems to take account of the rejects' characteristic vectors. The key assumption of the method is that $P(g|\mathbf{x})$ is the same for the accepts and the rejects.

The first step is to develop a model which discriminates between the accept and reject groups instead of the goods and bads. The same selection and weighting techniques that will be used to build the final scorecard are employed. For

each score under this initial model, the proportion of accepts in the full sample are calculated. This acts as a probability of acceptance for each member of the accepts sample. The accepts are weighted proportional to the inverse probabilities of acceptance. The weighted accept sample is then used to construct a scorecard that discriminates between goods and bads in the normal way.

We have two criticisms of this approach:

(1) The augmentation method is not applicable if all the characteristics used to make the original accept/reject decision are available for inclusion in the new scorecard. If this is the case then we can achieve perfect discrimination between the accept and reject classes by employing the original classifier. There would then be no regions of the new characteristic space where the accept and reject classes overlap. As a result it would not be possible to reweight the good/bad probabilities for the accepts to take account of the characteristic vectors for the rejects (as the augmentation method attempts to do).

On the other hand, if the set of characteristics Y available for building a new scorecard is a subset of the set X used in the original classifier, then we know from Sections 6.2.3 and 6.2.4 that the accept sample will be biased. The probability of being good amongst the accepts at any point in the characteristic space may not equal the corresponding probability for the rejects. Thus, when the accepts are reweighted to take account of rejects with similar characteristic vectors bias will be introduced.

(2) If an accurately specified model is used, then the weighting described above will have no effect on the estimated parameters. Therefore, augmentation does not offer any objective advantages over models built on the accept sample. As we shall show in Section 6.5.3, the characteristic vectors for the rejects cannot be used to improve a scorecard, except by chance or if additional information is being incorporated in some way.

Method 2

Joanes (1993/4) proposes an approach to reject inference which utilises iterative reclassification, a technique described in the context of discriminant analysis by McLachlan (1975).

The proposed method involves the construction of a classification rule using a logistic regression model estimated from the accepts sample. The logistic regression model can be expressed by the log-odds ratio:

$$\text{logit}(P) = \log \frac{P(g|\mathbf{x})}{P(b|\mathbf{x})} = \beta_0 + \beta_1 x_1 + \ldots \beta_p x_p .$$

The classification rule can be expressed as:

Accept applicant if $\text{logit}(P) > 0$ and

Reject applicant if $\text{logit}(P) \leq 0$.

The author states that it is necessary to incorporate the prior probabilities of class membership and the relative costs of misclassification into the classification rule. Given prior probabilities π_g and π_b and costs of misclassification $c(g|b)$ and $c(b|g)$, the new rule is defined to be:

Accept applicant if $\beta' \mathbf{x} + \log \frac{n_b \pi_g}{n_g \pi_b} \leq \log \frac{c(g|b)}{c(b|g)}$

and reject applicant otherwise.

Joanes proposes two methods of iteratively reclassifying the rejects in order to produce an unbiased scorecard: the first (method A) involves allocating the rejects to the class to which they have the highest probability of belonging (based on the accept model) and the second (method B) involves allocating class membership probabilities to the rejects from the accept model. In both cases a new classification rule is then constructed using the accepts with their true good/bad classification and the rejects with their estimated class or good/bad probabilities. The process is iterated until there is no change in the predicted class or probabilities of the rejects. This approach is a form of extrapolation as described in Section 6.5.1.

In order to illustrate how the method works, it is applied to a simulated data set with three characteristics. It is shown that the first and second iterative

reclassification procedures converge after two and six iterations respectively. These two approaches to reject inference have several limitations:

- There is no objective basis for classifying rejects as good or bad at each stage in method A. The method only serves to reinforce the structure of the accepts model. This can be seen from looking at the adjusted posterior probabilites which only show a marginal change from the accepts model to the model after the first iteration. Furthermore, there is no change in the adjusted posterior probabilities in subsequent iterations.

- If the prior probabilities and relative costs of misclassification were excluded from the classification rule then the iterated models from method B would be identical to the accepts model. (This is shown for linear regression in Section 6.5.1.) The only difference in the model comes from incorporating extra (subjective) information (and this information makes little difference to the posterior probabilities).

- We show in Section 6.5.3 that the characteristic vectors for the rejects do not contain information about the model parameters. Therefore, neither method A or B is able to reduce the bias of the accepts scorecard. These approaches could only be of use if meaningful values are known for the relative costs of misclassification.

A more general criticism of Joanes's paper is that no attempt is made to compare the performance of methods A and B with other reject inference methods, such as extrapolation. (This was partly due to the unavailability of a real data set.) We conclude that iterative reclassification is not a suitable approach to reject inference.

Method 3

Reichert et al. (1983) discuss a different way of tackling reject inference without incorporating extra information. They considered a three-group discriminant model (the groups were goods, bads and rejects) and compared it with a two-group (goods, bads) model based on the accepts sample. They found that the introduction of the rejects did not lead to improved

discrimination between the known goods and bads. It simply split the bads into two: half were predicted to be rejects and the other half to bads. However, the three-group model was assumed to be superior because it was conceptually more appropriate than a two-group truncated model in that it applied to the entire applicant population. Reichert et al. (1983) stated that this was also desirable because it accorded with regulatory rulings pursuant to the U.S. Equal Credit Opportunity Act which require that statistical credit scoring systems represent the complete population of applicants for credit.

Despite these apparent advantages the method adopted is subject to criticism. The assertion (on page 105) that the "key decision is whether to grant credit in the first place and not to identify good or bad borrowers once a loan has been granted" is true, but it fails to take into account that the main objective is to split the population of applicants into two groups. Therefore it does not seem appropriate to use a three group model. We can identify the rejects' class perfectly using the original classifier and so it seems odd to attempt to incorporate this function into a new classifier. Some form of extrapolation approach would be more appropriate.

One other potential advantage of this method arises with classical linear discriminant analysis. The method assumes a common covariance matrix for the three (or however many) groups in the data. The rejects can lead to more accurate estimates of this common covariance matrix and as such may lead to slight improvements in the classification rule. However, it is unlikely that this factor will outweigh the negative aspects of the method. Furthermore, in Section 3.1.1 we saw that credit scoring data often do not satisfy the assumptions of linear discriminant analysis. In many real situations the sample covariance matrices are unequal leading to the rejection of a linear discriminant rule in favour of a quadratic rule. Method 3 does not give the same benefit in this case.

6.4. Extrapolation from the accepts

A fundamental approach to reject inference is to build a model using the accepts sample, with their known characteristic vectors and true creditworthiness, and extrapolate it over the reject region. In Section 6.2 we considered the performance of a simple extrapolation model and identified aspects of the original classifier which relate to bias in the model: the extent of the truncation (the size of the reject region); the discriminatory power of the original classifier and the relationship between the characteristics in the original classifier and those available for constructing new scorecards. We found that in some cases extrapolation can provide surprisingly good performance. However, we restricted attention to building credit scoring models using linear regression and representing the data using weights of evidence. Although the factors identified are important in explaining bias, they do not provide a means of choosing between different classification techniques and data transformations for extrapolation puposes. This motivates a more general theoretical discussion of the effect of data structure and classifer on the performance of extrapolation models.

6.4.1 Theoretical aspects of extrapolation from the accepts

There are three basic features of the data that may cause a scorecard built on the accepts to extrapolate badly:

(1) If the required model form for the data differs between the accept and reject regions. For example, if the data has a linear relationship with creditworthiness in the accept region and a quadratic relationship in the reject region. In this case it is not possible to reject the hypothesis of a linear predictor in favour of a quadratic predictor from the accept region. Extrapolation will produce less and less accurate results as the departure from linearity becomes more extreme.

(2) If there is too much variation in the accept region for the true relationship between creditworthiness and score to be determined. In this case sampling variation will make it hard to identify higher order relationships.

(3) If an incorrectly specified model is used in the accept region. This is obviously more a property of poor model construction than a feature of the data. However, it is usual for scorecards to be constructed using standard methods, such as linear or logistic regression, without much consideration of whether the data really fits the model. Therefore, if the data differs from this "standard" model then we can consider it as a property of the data.

If the data structure falls into any of the three categories described above then a model built on the accepts will probably perform significantly less well than a model built on the full sample. This means that before deciding to carry out a particular form of reject inference careful examination should be made of the data. However, we believe that credit scoring data is often monotonic, if not quite linear, so that the above conditions are unlikely to be met.

The other factor that can have an important effect on the performance of extrapolation from the accepts is the type of model used for classifier design. We define, as in Chapter 2, the probability density function for the good risk applicants to be $P(\mathbf{x}|g)$, with $P(\mathbf{x}|b)$ the corresponding function for the bads. There are two possible approaches that can be taken to building a scorecard on the accepts (the diagnostic and sampling approaches of Dawid (1976)). The first involves direct estimation of the $P(g|\mathbf{x})$ and includes techniques like linear and logistic regression. The second approach is to estimate the $P(\mathbf{x}|g)$ and $P(\mathbf{x}|b)$ separately and then use Bayes theorem to calculate $P(g|\mathbf{x})$; an example of this type of method is classical linear discriminant analysis.

Our assertion is that there is an important difference between these two approaches when used for building a model on the accepts. Direct methods of estimating $P(g|\mathbf{x})$ are less prone to introducing bias into the classifier than indirect methods. This is because the direct methods are less likely to give distorted models for $P(g|\mathbf{x})$ when sampling fractions are taken from the characteristic space, that are only dependent on $\mathbf{x}$. In our context the accept/reject decision is based solely upon the characteristic vectors $\mathbf{x}$ and leads to a zero sampling fraction in the reject region and normal sampling in the accept region. This means that, subject to the conditions mentioned earlier in this section, we would expect models built on the accepts using direct methods to perform well and those constructed using indirect methods to perform less well. Two simple examples are presented to illustrate the point:

(1) A direct method of classifier design: assume that the data is distributed about a straight line along the attributes of one characteristic that is to be used to build a classifier. If a linear regression model is built on the accept sample then it will perform well over the reject region as illustrated in Figure 6.8.

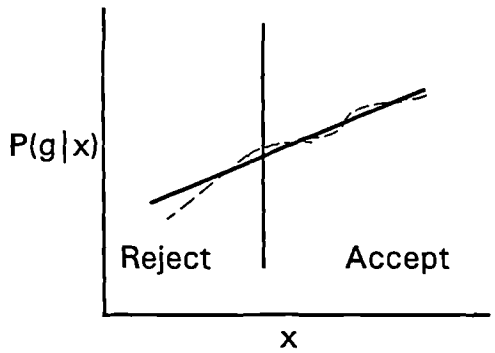


Figure 6.8: An example where a direct method of estimating $P(g|\mathbf{x})$ can lead to a good model when only the accept sample is used.

The curve represents the true distribution of the data and the line represents the fitted regression line.

(2) An indirect method of classifier design: assume that the $P(\mathbf{x}|g)$ and $P(\mathbf{x}|b)$ are normally distributed for one characteristic that is to be used to build a model to predict creditworthiness. Figure 6.9 shows the true distributions of the goods and bads. The curve for the bads distribution in the reject region is represented by a dotted line.

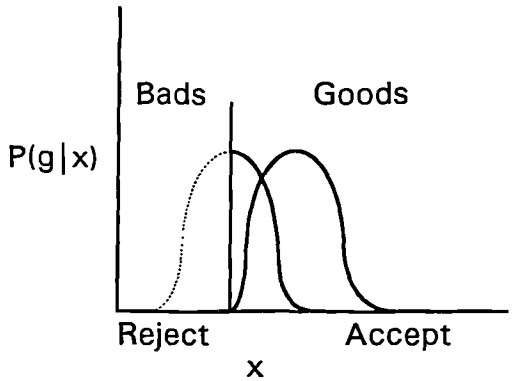


Figure 6.9: An example where an indirect method of estimating $P(g|\mathbf{x})$ can lead to a biased accepts model.

Using just the accepts to estimate the group means and dispersion matrices leads to distorted results for the $P(\mathbf{x}|b)$ distribution. Eisenbeis (1978) describes how Avery (1977) has looked at the effects of estimating discriminant analysis models for truncated sub-samples from the full distributions of goods and bads, where the full distributions are normal with known population parameters. He showed that one obtains biased estimates of the group means and dispersions, the true cut-off points and the true error rates. The direction and extent of the bias depends on the original truncation points and so is not readily available to the developer of a credit scoring model. Furthermore, the results show that even when the underlying populations have equal dispersions, the use of truncated samples leads to the rejection of the equal dispersion hypothesis. This means that quadratic rules will be selected in favour of linear ones.

To summarise the above discussion, we have investigated the factors that affect whether extrapolation from the accepts can lead to adequate models for $P(g|\mathbf{x})$. The main factors that have been identified are the type of model used to make the accept/reject classification, the nature of the data, the adequacy of the model chosen to represent the data and whether it models $P(g|\mathbf{x})$ directly or indirectly. In cases where the accept/reject decision is based solely on the characteristic vectors $\mathbf{x}$, we would expect direct models such as logistic regression to extrapolate quite well. This was confirmed by the second example presented in Section 6.2 (but not by the first example). In contrast we would expect that indirect methods of model building, such as classical linear discriminant analysis, would not perform well when used to build a model on the accepts. It might be possible to adjust the model to take account of the truncated sample, but the success of this approach would depend on how accurate the model assumptions were (e.g. the multivariate normality assumptions).

Extrapolation from the accepts is an intuitively simple and appealing approach to building a scorecard when the true creditworthiness of rejected applicants is not known. When this approach is used, a direct method of estimating $P(g|\mathbf{x})$ should be employed to build the scorecard. The results obtained will form a natural baseline against which to compare more sophisticated methods of reject inference that will be considered later.

6.5 Using the characteristic vectors for the rejects

Much attention has been focused on using the rejects to improve on simple extrapolation. In Section 6.3 we discussed methods of reject inference which have been proposed in the literature: augmentation (Hsia, 1978), iterative reclassification (Joanes, 1993/4) and a three group discriminant model (Reichert et al., 1983). It was found that all the proposed methods have theoretical weaknesses.

In this section we show that the rejects do not contain information that can be used to improve the predictive power of a scoring model, unless the new rule is better than the old one by chance or if expert knowledge is being incorporated in some form. This will be approached from several angles in Sections 6.5.1 to 6.5.3. In section 6.5.4 we will discuss how the assumption of particular forms for the the separate distributions of the characteristic values of the goods and bads, $P(\mathbf{x}|g)$ and $P(\mathbf{x}|b)$, can allow $P(g|\mathbf{x})$ to be calculated by mixture decomposition methods. This is one way in which the characteristic values for the rejects can contribute to the scorecard, but it is implicitly making use of extra information through the assumption of distributional forms for $P(\mathbf{x}|g)$ and $P(\mathbf{x}|b)$. More extensive discussion of reject inference methods which incorporate extra information is provided in Section 6.6.

6.5.1. Use of extrapolation

To begin with we will look at simple ways of using the characteristic vectors for the rejects as part of a reject inference strategy. One approach is to build a model on the accept sample and extrapolate from this model to provide estimates of $P(g|\mathbf{x})$ for the rejects. A new model could then be constructed using the accept sample with their true good/bad classification and the rejects with their estimated creditworthiness. This may be an appealing idea but there will be no change in the final scorecard if the same method of model building is used for constructing the accept model and the final model. This is easy to show and we will do so below in the special case of linear regression. Two approaches are described: an algebraic proof and an argument based upon residuals.

Approach 1:

Let X_a and X_r represent the data matrices for the accept and reject data respectively, with corresponding class vectors Y_a and Y_r. The values of Y_r are unknown and so are estimated from a model built on the accepts. A linear regression model for the accepts is given by

$$E(Y_a) = X_a\beta_a \quad (6.1)$$

where β_a is the vector of parameters to be estimated.

The parameters are estimated by maximum likelihood estimation giving (see for example Draper and Smith, 1981)

$$\hat{\beta}_a = (X_a^t X_a)^{-1} X_a^t Y_a \quad (6.2)$$

Estimates of the true classes for the rejects are then obtained by extrapolation from the accepts model to give

$$\hat{Y}_r = X_r\hat{\beta}_a \quad (6.3)$$

The final scorecard is obtained by building a model using the accept sample with their true classes and the rejects with their estimated classes. Let Y represent the full vector of classes given by $Y = (Y_a^t, \hat{Y}_r^t)^t$ and X the corresponding data matrix given by $X = (X_a^t, X_r^t)^t$. The parameters of the full sample model, β, are then given by:

$$\begin{aligned}
\beta &= (X^tX)^{-1}X^tY \\
&= [(X_a^t, X_r^t)(X_a^t, X_r^t)^t]^{-1}(X_a^t, X_r^t)(Y_a^t, (X_r\hat{\beta}_a)^t)^t \\
&= [X_a^tX_a + X_r^tX_r]^{-1}(X_a^t, X_r^t)(Y_a^t, (X_r\beta_a)^t)^t \\
&= [X_a^tX_a + X_r^tX_r]^{-1}[X_a^tY_a + (X_r^tX_r)\beta_a] \\
&= [X_a^tX_a + X_r^tX_r]^{-1}[X_a^tX_a + X_r^tX_r]\beta_a \qquad \text{by rearranging 6.2} \\
&= \beta_a \ .
\end{aligned}$$

We have shown above that the new full sample model will have the same parameter estimates as the accept sample model, and thus the characteristic vectors (in X_r) are not adding information to the model. Therefore this approach to reject inference will not lead to an improved scoring rule. Similar results can be proved for generalized linear models.

Approach 2:

Another way to obtain the same result for linear regression is to consider residuals. Linear regression estimates parameters by choosing them so as to minimise the overall squared differences between observed values and fitted values. This is equivalent to minimising the sum of squared residuals. If the

true classes of the rejects are estimated by $\hat{Y}_r = X_r\hat{\beta}_a$, then the rejects have zero residuals for the accepts model. Furthermore, the parameters of the accepts model were chosen to minimise the residuals for the accepts sample. Thus, if the accepts sample, with true creditworthiness known, is combined with the rejects sample, with true creditworthiness estimated from the accepts model, the parameters which minimise the overall sum of residuals will be equal to the parameters of the accepts model. As above, the characteristic vectors for the rejects do not add information about the full model parameters.

The previous discussion indicates that the rejects' characteristic vectors alone are not sufficient to reduce the bias of models built on the accepts sample. However, if one is prepared to make additional assumptions, then it may be possible to incorporate successfully the characteristic vectors for the rejects into an extrapolation approach. We consider two possible approaches:

(1) In Section 6.4.1 we discussed the situation where the required model form differs between the accept and reject regions. We considered a one dimensional example where the data has a linear relationship with creditworthiness in the accept region and a quadratic relationship in the reject region. If a linear model is built on the accepts and extrapolated over the reject region then biased estimates of $P(g|\mathbf{x})$ will be obtained for the rejects. Knowledge of the shape of $P(g|\mathbf{x})$ in the reject region could be obtained using a calibration sample (see Section 6.6). One possible extrapolation approach, which uses this additional information, is to build a model on the accepts using the appropriate form for extrapolation over the reject region (a quadratic model in the above example). The bias in the estimates of the true creditworthiness for the rejects can thus be reduced.

(2) Another possible approach to using the characteristic vectors for the rejects is to make the assumption that the rejects with the lowest scores under the original accept/reject classifier (if available) have a zero probability of being good ($P(g|\mathbf{x}) = 0$). A model can then be built using the accepts with their true creditworthiness and the worst rejects with $P(g|\mathbf{x})=0$ (a constrained regression) and this can then be extrapolated over the reject region to provide estimates of $P(g|\mathbf{x})$ for the remaining rejects. The validity of the initial assumption can be assessed using a calibration sample. Examination of our sample confirmed that

it is a reasonable assumption to make for a small proportion of the rejects. However, this approach is ad hoc and too dependent upon the original classifier.

In both of the above examples the extra assumptions will lead to a difference between the accept and full sample models, and possibly to an improvement in discrimination.

6.5.2 Standard missing data approaches

One way of looking at reject inference is to regard it as a missing data problem. This area has been the subject of much recent study in statistics, with the increased sophistication of computers and statistical packages allowing practical application of techniques often requiring intensive calculations. A description of many of these techniques is given in Little and Rubin (1987). Four basic procedures are considered:-

(a) *Removing incompletely recorded observations and analyzing the complete data.*
In our terms this means working just with the accepts. This approach has been covered in detail in Sections 6.2 and 6.4.

(b) *Imputation-based procedures.*
Missing values are filled in with values estimated from the observed data (e.g using the class of the "nearest" point in the accept region- cf. *k*-NN method in Chapter 8) and the resultant data set is then analysed using complete-data techniques. This approach differs from extrapolation and approach (d) (see below) in that it does not use a model for the observed data to estimate the missing values

(c) *Weighting procedures.*
This approach is used for randomization inferences from sample survey data with nonresponse. For example, the population mean can be adjusted for nonresponse by weighting every term by an estimate of the corresponding probability of response. (Weighting is related to mean imputation.)

(d) *Model-based procedures.*

This heading covers a wide range of techniques that involve defining models for the partially missing data and basing inferences on the likelihood under that model. The likelihood for the parameters based on the incomplete data is derived and maximum likelihood estimates are found by solving the likelihood equation. By assuming a model form in this way we are assuming extra information (as in the mixture decomposition approach). This approach is more flexible and less ad hoc than the first three procedures and was the starting point for our investigation of possible methods of reject inference. The extrapolation approaches we have considered in previous sections can also be described as model-based procedures.

Two important components of the work on model-based procedures are the consideration of different missing data mechanisms and the solution of the likelihood equations using the EM algorithm. These are both considered in detail below.

6.5.2.1 Missing data mechanisms

Little and Rubin (1987) point out that knowledge, or absence of knowledge, of the mechanisms that led to certain values being missing is a key element in choosing an appropriate analysis and in interpreting the results. The mechanism that gives rise to missing data corresponds to the method used to make the accept/reject decision in the credit scoring problem. In what follows we assume that there are more than one observable variables (characteristics), but only one variable Y (creditworthiness) is subject to nonresponse. Little and Rubin describe three mechanisms for missing data which have different implications for maximising the observed data likelihood (see Section 6.5.3).

(1) The data is *missing completely at random* (MCAR) if the probability of response is independent of all the variables.

(2) The data is *missing at random* (MAR) if the probability of response does not depend on Y, but may depend on one or more of the other variables

(3) The data is *non-ignorably* missing (NI) if the probability of response depends on Y and one or more of the other variables.

In case (3) the missing data mechanism is *non-ignorable*, meaning that in making likelihood-based inferences it is necessary to model the missing data mechanism. In both cases (1) and (2) the missing data mechanism is ignorable for likelihood-based inferences. Whether the mechanism is ignorable or non-ignorable has significant implications for maximising the observed data likelihood.

In the credit scoring problem, the appropriate missing data mechanism depends upon the set of characteristics used to make the original accept/reject decision. We distinguish between two possible scenarios as in Section 6.2.3:

(1) If the set of characteristics Z available for building new scorecards includes all the characteristics from the set X used to make the original accept/reject classification then the data will be MAR. This is because the probability of response depends on the set of characteristics Z alone and not on the response Y.

(2) If the set of characteristics Z available for building new scorecards does not include all the characteristics from the set X used to make the original accept/reject classification then the data will be NI. This is because the probability of response is indirectly dependent upon the true creditworthiness Y through the characteristics in X / Z.

In Section 6.5.3 we present a likelihood approach to incorporating the characteristic vectors for the rejects into a scorecard. We consider different non-response models for the three missing data mechanisms. In Hand and Henley (1994) we apply our arguments in a medical screening context.

6.5.2.2 The EM algorithm

With some patterns of missing data it is possible to use standard complete-data techniques to solve directly the likelihood equations (see Section 6.5.3). However, in practical problems it is often the case that closed-form solutions of

the likelihood equations cannot be found and so an iterative technique needs to be employed. The most widely used technique in the statistical literature is the *Expectation-Maximization* (EM) algorithm which relates maximum likelihood (ML) estimates for the observed data likelihood to ML estimation based on the complete-data likelihood (including missing values). It was introduced by Dempster et al. (1977). Because of its central importance in the field of missing data problems it was thought necessary to examine the EM algorithm and determine its applicability to our reject inference problem. A second reason for considering the EM algorithm was suggested because of the striking similarity between the iterative nature of the mail order company's current method of attempting to improve scorecards using reject inference and the iterative steps of the EM algorithm.

In fact, the EM algorithm is only an iterative method for solving the likelihood equations, based on the observed data, when a closed form solution does not exist. In the special cases where the rejects true status is estimated using equation (6.1), the EM algorithm converges to the same parameter values that are obtained from a model built on the accept sample (the proof of this is analogous to approach 1 in Section 6.5.1). This means that the rejects do not provide any information about class membership. This result is stated in Section 8.4.1 of Little and Rubin (1987):

"When a scalar outcome variable Y is regressed on p predictor variables $X_1, X_2 \ldots X_p$ and missing values are confined to Y, the incomplete observations do not contain information about the regression parameters."

The result of this is that although the EM algorithm can be applied to our problem of reject inference, it does not convey any advantage over simply constructing a scorecard using the original accept sample. It is worth noting that the EM algorithm could be useful in tackling a slightly different problem: namely how to perform maximum likelihood estimation of model parameters when characteristic information is missing for some of the applicants in the sample. This is likely to be the case in practice, but we choose to avoid the problem by creating an attribute called "no information".

6.5.3 A likelihood based approach

In this section we provide a likelihood based argument to show that the characteristic vectors for the rejects do not provide information about the parameters of the observed data model, except when certain specific conditions are satisfied. It should serve to clarify our view of reject inference in a rigorous manner.

We begin by defining the terms we shall use.

Let Y_a be the observed good/bad indicator for the accepts,
Y_r be the unobserved good/bad indicator for the rejects,
X_a be the observed characteristic matrix for the accepts,
X_r be the observed characteristic matrix for the rejects,
b be the observed accept/reject indicator (0 for reject, 1 for accept),
and $\alpha,\beta,\gamma,\delta$ be parameter vectors for the indicated probability functions. (We shall write all such functions as $f(\cdot)$, relying on their arguments to distinguish them.)

The overall likelihood of the observed data is then given by:

$$\begin{aligned} L &= f(Y_a,X_a,X_r,b;\alpha,\beta,\gamma,\delta) \\ &= \int f(Y_a,Y_r,X_a,X_r,b;\alpha,\beta,\gamma,\delta)\,dY_r \\ &= \int f(b|Y_a,Y_r,X_a,X_r;\alpha) f(Y_a,Y_r,X_a,X_r;\beta,\gamma,\delta)\,dY_r \end{aligned} \qquad (6.5.1)$$

The second term inside the integral in (6.5.1) is the full data model, which describes how the data arises. The first term is the non-response model. It corresponds to the classifier used to make the original accept/reject decision. The parameter vector α can be regarded as constant, because it is fixed by the original classifier. Despite this, the non-response model needs to be considered because it affects estimation of the parameters in the data model through the integration of (6.5.1).

We consider the non-response models corresponding to the three missing data mechanisms described in Section 6.5.2.1.

(1) If the data are MCAR the non-response model factorises to give:

$$f(b|Y_a,Y_r,X_a,X_r;\alpha) = f(b;\alpha)$$

(2) If the data are MAR the non-response model factorises to give:

$$f(b|Y_a,Y_r,X_a,X_r;\alpha) = f(b|X_a,X_r;\alpha)$$

(3) If the data are NI the non-response model does not factorise.

The MCAR case is not of further interest because it corresponds to the original accept/reject decision being made independently of the rejects' characteristic values.

The NI case corresponds to the new classifier being constructed on a set of characteristics which does not include all of the set used to make the original classification. In this situation the observed data likelihood cannot be factorised. It is not possible to obtain explicit maximum likelihood estimates for the model parameters. In particular, the problem does not reduce to estimating the parameters for the accepts data. Thus, a model built on the accepts sample will be biased. This justifies the conclusion of Section 6.2.2.

The most common missing data mechanism for credit scoring data is MAR. This occurs when all the characteristics used for making the original selection are available for constructing the new scorecard. In this case the observed data likelihood can be factorised to give:

$$\begin{aligned} L &= f(b|X_a,X_r;\alpha)\int f(Y_a,Y_r,X_a,X_r;\beta,\gamma,\delta)dY_r \\ &= f(b|X_a,X_r;\alpha)f(Y_a,X_a,X_r;\beta,\gamma,\delta) \\ &= f(b|X_a,X_r;\alpha)f(Y_a,X_a;\beta,\gamma)f(X_r;\delta) \quad \text{by independence} \\ &= f(b|X_a,X_r;\alpha)f(Y_a|X_a;\beta)f(X_a;\gamma)f(X_r;\delta) \qquad (6.5.2) \end{aligned}$$

Given the data is MAR, we consider the implications of different assumptions about the sampling scheme:

(1) When the sampling is conditional on x the functions $f(X_a;\gamma)$ and $f(X_r;\delta)$ are prespecified and so can be ignored for parameter estimation. As mentioned above, the parameter vector α comes from the classifier used to make the original accept/reject classification. Therefore, the term $f(b|X_a,X_r;\alpha)$ can be disregarded and the likelihood reduces to

$$L \propto f(Y_a|X_a;\beta)$$

This means that estimating the parameters of the observed data likelihood reduces to estimating the parameters, β, of the model for the accepts data. This means that including the characteristic vectors for the rejects does not lead to an improved scoring instrument.

(2) When sampling is from the overall mixture distribution and there is no intersection between the parameter spaces for $\alpha,\beta,\gamma,\delta$ then the observed data likelihood reduces to

$$\mathrm{L} \propto f(Y_a|X_a;\beta)$$

as in the above case. As before the characteristic vectors for the rejects do not contribute to the parameter estimates. This sort of scenario is common in practice and can occur when direct methods of estimating $P(g|\mathbf{x})$, such as logistic regression, are used with no assumptions made about the separate distributions of goods and bads.

On the other hand, if we make assumptions about the distributional form of the probability density functions for the good and bad applicants, $P(\mathbf{x}|g)$ and $P(\mathbf{x}|b)$, then the parameter space spanned by β intersects with the parameter space spanned by γ and δ. In this case the characteristic values for the rejects contribute to the parameter estimation. This apparent gain in information comes from the assumption of distributional forms for the goods and bads separately. The reasonableness of these assumptions will determine the success of this approach to reject inference. This mixture decomposition approach is discussed in more detail in Section 6.5.4.

Our conclusion from this section, and one that has important implications for developers of credit scoring systems, is that reliable reject inference is impossible unless additional information is available or assumptions are made.

6.5.4 Method 4: the mixture decomposition approach to reject inference

If one makes assumptions about the distributions of goods and bads separately, $P(\mathbf{x}|g)$ and $P(\mathbf{x}|b)$, and an indirect method of classifier design is used, then it is possible to include the rejects in the parameter estimation to reduce bias of models built on the accepts.

As an example of this approach, suppose that one is using linear discriminant analysis to build credit scoring models. Thus, we can assume that $P(\mathbf{x}|g)$ and $P(\mathbf{x}|b)$ are normally distributed with population means μ_1 and μ_2, and common covariance matrix Σ.

In the MAR case, the likelihood function is given by:

$$L = f(b|X_a,X_r;\alpha)f(Y_a|X_a;\beta)f(X_a;\gamma)f(X_r;\delta)$$
$$= f(b|X_a,X_r;\alpha)f(Y_a|X_a;\beta)f(X;\gamma,\delta)$$

where $f(X;\gamma,\delta)$ is the overall distribution of applicants (for accepts and rejects).

We begin by showing that, as mentioned in the previous section, the parameter space spanned by γ and δ intersects with the parameter space spanned by β. To see this we consider the two parameter spaces separately.

First, the parameter estimates of the discriminant function are:

$$\hat{\mathbf{a}} = \Sigma^{-1}(\mu_1 - \mu_2) \qquad (6.5.3)$$

If we incorporate prior probabilities $\mathbf{p} = \{P(g),P(b)\}$ into the discriminant function then the parameter vector $\beta = \{\Sigma^{-1}(\mu_1 - \mu_2),\mathbf{p}\}$.

Secondly, we consider the parameter space for the overall distribution of characteristics. Ignoring the parameter vectors, the appropriate distribution function can be expressed as:

$$f(X;\gamma,\delta) = f(\mathbf{x}) = P(\mathbf{x}|g).P(g) + P(\mathbf{x}|b).P(b) \qquad (6.5.4).$$

Using this relationship we can express the parameter vectors γ and δ in terms of the parameters μ_1, μ_2, Σ and $\mathbf{p}$. Thus, the parameter vectors γ and δ contain information about $\beta = \{\Sigma^{-1}(\mu_1 - \mu_2),\mathbf{p}\}$.

Equation (6.5.4) gives a means of incorporating the characteristic vectors for the rejects into the model to reduce the bias. The function $f(\mathbf{x})$ on the left hand side of the equation can be estimated from the overall distribution of the sample. Then, given assumed values for the population priors $P(g)$ and $P(b)$, the parameters of the $P(\mathbf{x}|g)$ and $P(\mathbf{x}|b)$ distributions can be estimated using the EM algorithm. They are chosen so as to minimise the difference between the left and right hand sides of the equation. The parameters of these distributions can then be used in the discriminant function (6.5.3). In this way

the characteristic vectors for the accepts and rejects can be combined to estimate the model parameters. For this approach to be valid, applicants must be sampled randomly from the overall mixture distribution. If this is not the case then the parameter estimates will be invalid (although if the sampling mechanism can be described, it is possible to allow for it in the analysis).

Because the space spanned by β is a subspace of that spanned by γ and δ, one can find estimates of β (and hence $P(g|\mathbf{x})$) without knowing the true classes for any of the applicants in the sample. Therefore, given assumed distributions for $P(\mathbf{x}|g)$ and $P(\mathbf{x}|b)$, it is possible to estimate $P(g|\mathbf{x})$ throughout the characteristic space using just the rejected applicants.

The mixture decomposition method is an appealing approach to reject inference which attempts to make use of all available information. However, its relative performance is dependent upon the reasonableness of the assumed parametric forms for the data. Unfortunately, the assumption of multivariate normality may not be realistic for credit scoring data. Furthermore, there does not seem to be an appropriate parametric alternative.

One other disadvantage of the method (as described above) is that it is only compatible with indirect methods of estimating $P(g|\mathbf{x})$. We conclude this section by considering whether the approach can be adapted to apply directly to the $P(g|\mathbf{x})$ functions. This would enable its application when methods such as logistic regression are used for building scorecards.

The equivalent expression to equation (6.5.4) is

$$P(g) = \int P(g|\mathbf{x}) f(\mathbf{x}) d\mathbf{x}$$

The parametric form for $P(g|\mathbf{x})$ in the logistic case is

$$P(g|\mathbf{x}) = \frac{1}{(1+e^{-\beta'\mathbf{x}})},$$

where β is a vector of parameters.

In other words, we estimate β such that

$$P(g) = E\left(\frac{1}{(1+e^{-\beta'\mathbf{x}})}\right).$$

This problem is solvable in the special case when β consists of one parameter, but not when the data is multidimensional. Therefore, this approach cannot be applied to logistic regression in a general context.

6.6 Methods of reject inference that use supplementary information

We have seen that the characteristic vectors for the rejects do not provide useful information unless we are prepared to make extra assumptions, such as the distributional assumptions considered in the last section. We now consider other ways in which supplementary information can be incorporated into the reject inference procedure.

In simple terms the problem of reject inference arises because there are regions of the characteristic space where no information is available about true creditworthiness (namely the reject region). The ideas considered so far have mainly focused on building a model in the region we do have such information (the accept region) and extrapolating the model over the reject region in some way. This may perform well, but suffers if there are differences between the good/bad probability structure of the two regions. Another appealing approach that avoids this is to collect some information about the nature of the reject region. The easiest way to do this is to collect a subsample of rejected cases. The problem with doing this is that the proportion of bad applicants in the rejected region is likely to be higher than in the accept region (assuming the present method of screening is sensible) and each accepted bad applicant represents a financial loss. Therefore, for this strategy to be profitable, the saving made from the reduction of bad debt due to increased accuracy of the new scorecard needs to be greater than the losses due to the bad loans among the rejected applicants taken on.

For the rest of this section we will assume that supplementary information on the rejects is available in the form of a *calibration sample*. The definition of a calibration sample that we shall use is a subsample from a full population of accepts and rejects that includes the good/bad definitions for the rejects. It will be used to adjust the scoring instrument constructed on the accepts from a current sample.

We need to draw a distinction between two possible options that have different implications for the use of the calibration sample: first, the calibration sample may be randomly drawn from the same population as the new sample that we are using for model construction and, secondly, the calibration sample may come from some other source such as a sample collected in a previous year or from a different product or geographical location.

In the first case the most obvious policy would be to combine the accepts and calibration sample and use a direct method of estimating the $P(g|\mathbf{x})$ function, such as logistic regression (we should not use an indirect method for the reasons outlined in Section 6.4.1). In this situation the new accepts sample may add little to the information contained in the calibration sample (depending on the relative size of the two samples).

In the second case the above method may lead to a severely biased scorecard because the calibration sample will not give a representative picture of the population of interest. Therefore, a much more cautious use of the calibration sample is necessary. This could begin with an examination of the accepts portion of the calibration sample and a comparison with the new accepts sample to identify similarities of the two samples and how to incorporate these into a reject inference method. For the data that we were interested in a calibration sample was available that had been collected in time periods previous to the collection period for the current accepts sample. In the remainder of this section we will devote our attention to the consideration of this case where the calibration sample and new sample are from different populations. This is likely to be the case in most practical applications.

We consider four methods of reject inference that use calibration samples. The common theme is the assumption of various relationships between the calibration sample and the new sample. These relationships are used in two ways:

(1) To infer the performance of the rejects from the new sample and then construct a new scorecard using the accepts with their true creditworthiness and the rejects with their estimated good/bad probabilities

(2) To adjust a model built on the new accepts using the differences between the calibration sample accepts and the new sample.

6.6.1 Method 5

The first method that we consider here makes the least allowance for differences between the new and calibration samples. It is an intuitively simple procedure that has been developed and used in practice by the mail order company supplying the data.

The basis of the method is the construction of a scoring rule on the calibration sample and its subsequent use to infer good/bad probabilities for the rejects of the new sample. The new accepts with their true classification and the new rejects with their estimated good/bad probabilities are used to construct a new scoring rule. This is a way of using the calibration sample to provide a more informed extrapolation into the reject region.

The main weakness of the method is its dependence on the similarity of the new and calibration samples. Rejects with similar characteristic vectors will receive the same estimated good/bad probabilities regardless of the type of new sample that they come from. If the state of the economy had led to an overall reduction in the level of bad debt in the new sample, this would not be reflected in the estimated probabilities allocated to the rejects by the model from the calibration sample. This provides a motivation to look for more robust method of reject inference. Results of comparison with other methods will be presented in Section 6.7.

Variants of this method are possible that make some allowance for the differences between the two samples. One could provide good/bad probability estimates for the rejects using inferences from models built on both the calibration sample and the new accepts sample and combine these estimates in some way.

6.6.2 Method 6

This represents a new, and we believe original, approach to the problem of reject inference. It involves the partition of previous complete data sets (calibration samples) into sub-groups with particular properties. It is assumed that there are a significant number of these calibration samples and that between them they represent a variety of different good/bad probability structures that will be encountered in practice. The aim of the method is to model the relationship between $P(g|\mathbf{x})$ in the accept and reject regions for different samples. For a new sample with known class membership probabilities in the accept region, the method provides estimates of $P(g|\mathbf{x})$ in the reject region. We outline in general terms the various stages involved in the construction of reject inference models using Method 6.

The first stage of the method is the selection of suitable calibration samples. One approach is to divide a sample up into the individual years in which it was collected. This enables one to model how the relationship between creditworthiness in the accept and reject regions changes over time (for a discussion of population drift see Hand and Henley (1994)). Other ways of obtaining these multiple calibration samples are to collect samples from different geographical locations or for different products.

The second stage is to partition the calibration samples into accepts and rejects using the original classifier that was used to partition the new sample. (It is assumed that the original classifier is available along with the same range of characteristics for each sample.) Characterising features of the $P(g|\mathbf{x})$ distributions in both the accept and reject regions are then identified. This is possible because the calibration samples include the true classes for the rejects.

The third stage is to model the relationship between the characterising features of the accepts distribution and the rejects distribution. These relationships are then used to map from the characterising features of the new accepts sample to give predictions of the characterising features for the new sample rejects. These predictions of the characterising features map to give estimates of $P(g|\mathbf{x})$ for the new rejects. A classifier can then be constructed using the new accepts with their true class and the new rejects with their estimated $P(g|\mathbf{x})$.

6.6.2.1 A simplified application

We consider a simplified application of the general procedure outlined above in order to gain insight into the potential of the method for reject inference. It is assumed that there are few enough cells (combinations of attributes for all the characteristics in the model) to use the proportion good in a cell as the characterising feature of the accept and reject distributions. (In practice we will need to use more subtle summarising statistics, such as low order log-linear models. This issue is addressed in more detail in Section 6.6.2.2.). There follows a general outline of the method, given the above assumption, followed by the results of applying it to a simple data set.

Suppose that we start with n calibration samples. Each can be represented by a point in the space spanned by p_A and p_R where p_A is the vector of $p_i(good)$ values, with i ranging over the cells in the accept region, and p_R is the vector of $p_i(good)$ values, with i ranging over the cells in the reject region. In other words, we could plot a point for each calibration sample in r-dimensional space, where r is the total number of accept and reject cells. Figure 6.10 shows such a plot in the trivial case where there are only two cells: one reject cell and one accept cell. The vertical axis shows the predicted probability of an applicant in the reject cell being good and the horizontal axis shows the corresponding probability estimate for the accept cell. The "o"s represent different samples.

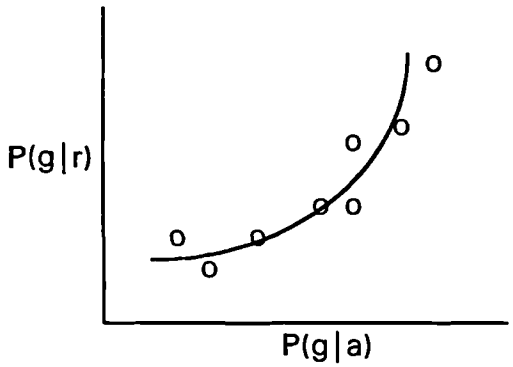


Figure 6.10: A hypothetical example of the relationship between $P(g|\mathbf{x})$ in the accept and reject regions for different calibration samples.

The next step is to fit a model (by a multivariate regression or a series of multiple regressions) to the n points in r-dimensional space. This gives the fitted curve shown in Figure 6.10. Using this model we can predict the good/bad probabilities for the reject cell(s) from just an accept sample. By doing this we obtain predicted probabilities of class membership for all the rejects in the design set. A new scorecard is then constructed using the accepts sample with their true status and the rejects with their predicted good/bad probabilities. We calculate a new acceptance region and could then iterate the process by obtaining updated good/bad probability estimates for those original rejects that are in the reject region of the new scorecard and so on.

This procedure was applied to a simple data set and a comparison was made with the results from extrapolation. The data set consisted of a calibration sample with 2189 applicants from two classes and a new sample with 3741 applicants again from two classes. (The new sample included the true status of the rejects so that we could compare our predictions with the true values). Although 26 characteristics were available, only two were used (each with three attributes) to keep the analysis simple. As a result there were nine possible cells of which eight were accept cells and one was a reject cell. The reject cell was selected under the criterion of lowest weight of evidence within each of the two characteristics used. In this example we have only considered one reject cell to avoid having to predict a vector of responses.

The calibration sample was split into ten sub-samples by splitting on one of the characteristics excluded from the analysis that related to geographical location. This gave us ten calibration samples with between 71 and 399 applicants in each one. These calibration samples were used to construct a linear regression model to predict the proportion good in the reject cell from the proportion good in the accept cells. This regression model was then applied to the accept cells of the new sample to give a prediction of the proportion good in the reject cell. Table 6.9 shows the results of Method 6 and extrapolation together with the true proportion good in the reject cell.

	$P(g \mid reject)$
True	0.27
Method 6	0.29
Extrapolation	0.1

Table 6.9: The predicted proportion good in the reject cell for extrapolation and method 6.

The results shown in Table 6.9 indicate that extrapolation severely underestimates P($g \mid reject$) for this set of characteristics. In contrast, Method 6 provides an accurate estimate of the true proportion good. The procedure was repeated several times using a different two characteristics selected from the 26 available each time. The results for Method 6 were found to be consistently closer to the true values than the extrapolation results. However, in some cases the results for Method 6 and extrapolation were very similar.

We conclude from this simplified analysis that information about the reject region contained in calibration samples can be used to provide more accurate estimates of the true proportion good. However, two outstanding issues remain to be resolved in order to allow practical implementation of the method.

First, the method requires access to enough samples to represent adequately the feature space for $P(g \mid accept)$ and $P(g \mid reject)$. Our calibration sample does not have natural sub-groups so it is necessary to split on characteristics which are not included in the analysis.

Secondly, there is very high dimensionality of both p_A and p_R. It is not possible to ignore the problem because there are 2^r cells for r binary characteristics and so the number of cells soon becomes unmanageable. One approach to solving the problem is to use shrinkage techniques. Another approach, which we adopt in Section 6.6.2.2, is to replace p_A and p_R by $\hat{p}_A$ and $\hat{p}_R$ predicted from a model and work with the parameters of the model. A similar approach would be to work with applicant scores (on the initial scorecard) rather than individual cells. This would be a way of grouping together cells with similar properties (the r-dimensions would be the total number of possible scores on the initial scorecard and this number can be controlled).

6.6.2.2 A second comparison of method 6 with simple extrapolation

We now address the second problem outlined above through a second comparison experiment of Method 6 with extrapolation. We consider the performance of the two reject inference methods for different numbers of characteristics (up to a maximum of six) and compare the accuracy of the predicted $P(g \mid \mathbf{x})$.

The data used in the analysis consisted of the calibration sample split into eleven sub-populations according to a characteristic, "media code", which was not used in the classifiers. "Media code" describes the media through which an applicant applied for credit (e.g. newspaper adverts, mailed offers). Six characteristics were selected for inclusion in the experiment by building a stepwise regression model. The characteristics were used in weights of attribute form and are shown below:

(1) A decision tree combining several characteristics related to previous credit history (14 attributes)
(2) A characteristic combining time at address with postcode (10 attributes)
(3) Weeks since last default with other retailers (7 attributes)
(4) Postcode groups (40 attributes)
(5) A characteristic related to previous mail order experience (4 attributes)
(6) Time on electoral role (6 attributes)

An independent classifier was used to identify the accept and reject regions of the different calibration samples. The rejects account for 22% of the full sample.

To avoid the problem of having too few applicants in each cell to estimate accurately the proportion of goods in that cell, it was decided to work with parameters of a model. To further simplify the analysis each of the reject parameters was estimated separately from the accept parameters rather than using a multivariate method. Specifically logistic regression models were constructed on the accepts and rejects for each of the eleven sub-populations in

the analysis sample. Each of the reject parameters was then regressed linearly on the accept parameters to provide a model predicting the reject parameters from the accept parameters. These estimates of the reject parameters can be used together with the weights of evidence to provide predicted good/bad rates for the reject cells/applicants.

The extrapolation approach involved building a scoring instrument on the accepts portion of the full calibration sample using logistic regression. Estimates of the good/bad probabilities for the rejects were then obtained by extrapolation.

Two independent test sets were used to validate the results: a hold-out sample from the same population as the calibration sample and a hold-out sample from the same population as the new sample. Performance was assessed by considering two measures of discriminability and one measure of reliability for applicants in the reject region (see Chapter 5). The first measure of performance was the average probability assigned to the true class given by:

$$A=1/n\sum \hat{P}(c_i|\mathbf{x}_i),$$

where $\mathbf{x}_i$ is the characteristic vector for applicant i and c_i is the true class of applicant i. A suitable measure of reliability, based on this discriminatory measure, is given by:

$$B=1/n\sum (2\hat{P}(c_i|\mathbf{x}_i)-1)\hat{P}(c_i|\mathbf{x}_i)$$

The second discriminability measure we consider is

$$C=1/n\sum (I(\mathbf{x}_i)-P(g|\mathbf{x}_i))$$ where $I(\mathbf{x}_i)$ is 1 if applicant i is good and 0 otherwise. A classification rule is achieving good discriminability if measure A has a high value and measure C has a low value. Reliability measure B tells us how accurately the system is predicting the true good/bad rates in the reject cells. A value close to zero indicates good reliability.

In this comparison experiment we place particular emphasis on the reliability of a classification rule. This is to gain insight into how accurately Method 6 can predict $P(g|\mathbf{x})$ for the rejects. The aim of this experiment is to explore the properties of Method 6 rather than provide a practical classification rule. (In practice our aim is to achieve maximum discriminability through minimising the bad rate amongst the accepts.)

The procedure described above was carried out for three, four, five and six of the selected characteristics in turn, and the three performance measures were calculated for the two validation samples in each case. The results are shown in Table 6.10. S1 refers to the hold-out sample from the full calibration sample with 839 rejects and S2 refers to the new sample with 102 rejects. The number in brackets refers to the number of characteristics included in the analysis. The column headings consist of the letter of the performance measure followed by an abbreviation of the method name in brackets ("M6" stands for Method 6 and "extra" for extrapolation).

	A (M6)	A (extra)	B (M6)	B (extra)	C (M6)	C (extra)
S1(3)	0.591	0.605	-0.007	-0.026	0.208	0.211
S2(3)	0.591	0.596	-0.019	-0.062	0.214	0.233
S1(4)	0.594	0.570	-0.016	-0.064	0.211	0.247
S2(4)	0.614	0.619	-0.046	-0.044	0.216	0.212
S1(5)	0.606	0.580	-0.022	-0.068	0.208	0.244
S2(5)	0.622	0.629	-0.084	-0.050	0.231	0.210
S1(6)	0.650	0.591	-0.114	-0.050	0.232	0.229
S2(6)	0.472	0.629	-0.101	-0.043	0.314	0.207

Table 6.10: The results of the second experiment to compare Method 6 with extrapolation.

Several comments can be made about the above results:

(1) The values of measure B remain approximately constant for the extrapolation method as the number of characteristics increase, whereas there is a sharp deterioration in reliability for Method 6. One reason for this deterioration is that as the number of dimensions increases, the number of predictors in the regression of the reject parameters on the accept parameters increases towards the number of cases (equal to the number of calibration samples). However, Method 6 has a lower reliability score than extrapolation when only three characteristics are used.

(2) The discriminability measures A and C show that Method 6 and extrapolation give similar discrimination between the goods and bads in the reject region across samples and numbers of characteristics. However, there is a significant drop in performance for Method 6 with 6 characteristics using Sample S2.

(3) This experiment has illustrated the difficulty of trying to pick suitable sub-populations to model the accept and reject feature spaces. The calibration sample was split by media type, which is one of the characteristics not usually used in the analysis, and although there was a range of different bad rate values across the various sub-populations, it was not possible to get an adequate representation of the relationship between the accept and the reject bad rates. This is borne out by the end results.

(4) The version of Method 6 described in this comparison did not involve a multivariate technique for predicting the reject parameters from the accept parameters. It is appropriate to use a regression which allows for multiple response variables in order to take into account the covariances between them. We carried out a further analysis using a canonical correlations analysis. However, the resulting canonical variates were non-significant. This is probably because the calibration samples do not give an adequte representation of the feature space as mentioned in (3). Furthermore, the number of observations (calibration samples) may be insufficient.

(5) Inaccurate β values estimated from a small calibration sample can detract from the effectiveness of the method. Therefore, it might be more appropriate to use a weighted linear regression to model the relationship between the parameter estimates for the accept and reject regions.

6.6.2.3 Conclusions

Despite its theoretical appeal, Method 6 is difficult to implement effectively. In particular we can identify several considerations that need further attention.

First, it is difficult to find ways of splitting the data into distinct populations that give a suitable representation of the feature spaces for $P(g|\mathbf{x})$ in the accept

and reject regions. If the data does not come in readily identifiable sub-populations then the best option appears to be to split on the attributes of a characteristic that is not needed in the analysis.

Related problems are the need to have sufficient numbers of applicants in each calibration sample to enable accurate estimation of the characterising features, and the need to have sufficient numbers of calibration samples to enable a suitable mapping to be set up. Further work is needed to identify a method of balancing these conflicting aims.

We have seen that Method 6 can perform well in low dimensions where the number of reject cells is small. However, as the number of dimensions increases performance deteriorates rapidly. More work is needed on the identification of characterising features in high dimensions. There are two main problems with using regression parameters (as in experiment 2): first, the individual parameters are not robust. This is because they do not represent the individual relationship of one characteristic with creditworthiness. Secondly, because the parameters represent the combined relationship between the predictor variables and creditworthiness, it is necessary to use a multivariate technique to estimate the relationship between the accept and reject parameters.

6.6.3 Method 7

We have seen that there are situations in which an extrapolation approach can provide a good scoring instrument for the full applicant population. The factors that effect this strategy were considered in Sections 6.2 and 6.4. However, the assumption that a model will extrapolate accurately into the reject region is risky unless some information is available about the structure of $P(g|\mathbf{x})$ in this region (such as a calibration sample). The characteristic space has been split into accept and regions because we believe there to be fundamental differences between them. Thus, there are questions as to the legitimacy of an extrapolation approach. This argument has been used to justify the approaches to reject inference described in the literature (see Section 6.3). We propose a new approach, which we believe to be original, that requires a single calibration sample. The aim of Method 7 is to use the information contained in

the calibration sample to adjust a model built on the new accepts to reflect the influence of the rejects.

As mentioned earlier, if the new and calibration samples are from the same population it is appropriate to take a "pseudo-Bayesian" approach to reject inference, where the calibration sample represents subjective information about the model parameters. It can be shown that this is equivalent to combining the two samples and constructing a scoring instrument on the enlarged sample (we refer to this as Method *ES* below). Method 7, like Method 6, is designed for the more general situation where the two samples may come from different populations. The one assumption that we make is that the relationship between $P(g|\mathbf{x})$ in the accept and reject regions will remain constant for different populations of credit applicants. Although similar in concept to Method 6, Method 7 does not require the existence of suitable calibration samples to span the set of characterising features. It also has an appealing and intuitive concept.

The first stage is to construct a scoring instrument on the new accept sample. Then the calibration sample is split into accepts and rejects using the classifier that was used to make the original accept/reject decision. We construct two scoring instruments on the calibration sample: one on the full sample and one on the accepts portion. These two models will differ to some extent due to the factors affecting the performance of extrapolation. Method 7 uses this difference to adjust the scoring instrument built on the new accept sample. Figure 6.11 shows the principle behind the method in the simplified case where only one characteristic is available. The relationship between the accept and full curves is the same for the calibration and new samples even though the appropriate model form changes from a straight line to a curve.

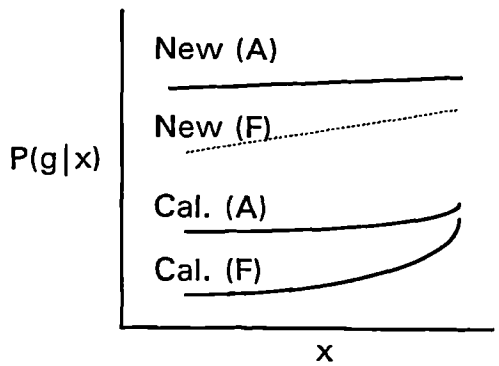


Figure 6.11: A hypothetical example where Method 7 can adjust the

model for $P(g|\mathbf{x})$ on the new accepts to take account of the difference between the full and accept calibration samples.

An important part of the method is the selection of an appropriate transformation to represent the difference between the models built on the full and accept calibration samples. We describe a simple first approach. Logistic or linear regression models are constructed for the new accept sample and the accept and full calibration samples. This gives sets of regression coefficients β_{NA}, β_{CA} and β_{CF}. The relationship between the coefficients for the calibration sample, β_{CA} and β_{CF}, can be described through a simple diagonal matrix **M** with entries given by the ratio of the β_{CA} to the β_{CF}. This matrix can then be used to give an estimate of the regression coefficients for the full new population, $\mathbf{M}\beta_{NA}$. (The regression coefficients serve a similar function to the characterising features of Method 6.)

Many other adjustment schemes are possible. Graphs of the β values for the full and accept calibration samples can be used to suggest suitable transformations. We summarise some simple transformations:

(a) $\beta_{CF} = \mathbf{M}\,\beta_{CA}$. The β_{NF} are estimated by $\beta_{NF} = \mathbf{M}\,\beta_{NA}$. In the simplest case **M** is a diagonal matrix of constants.

(b) $\beta_{NA} = \mathbf{N}\,\beta_{CA}$. The β_{NF} are estimated by $\beta_{NF} = \mathbf{N}\beta_{CF}$. In the simplest case **N** is a diagonal matrix of constants. It can be shown that (a) and (b) are equivalent if $\mathbf{MN} = \mathbf{NM}$.

(c) Fitting a simple one-dimensional regression model $\beta_{CF} = \alpha\,\beta_{CA}$ + error, where the values (β_{CF}, β_{CA}) are considered as observations from a bivariate sample.

We carried out a comparison experiment of Method 7 with Method *ES*. The samples used were a calibration sample (3102 cases with 2213 accepts) and a new sample (27000 cases with 18801 accepts). A crude classifier constructed on an independent sample was used to split the two samples into accepts and rejects.

Linear regression models were constructed on each of the relevant samples and the diagonal matrix **M**, as defined above, was calculated. This transformation was used to provide estimates of the linear regression coefficients for the new full sample. This adjusted model was then used to classify an independent test sample of 7189 applicants from the same population as the new sample. The analysis was carried out using 12 and 15 characteristics.

The range of samples/methods used to construct classifiers are described below:

(1) The full calibration sample and full new sample. This sample includes the reject data for the new sample (with true status known) and so is not a realistic option. It was included for completeness.

(2) The full calibration sample and the new sample accepts (Method *ES*).

(3) The full new sample. As with approach (1) this is not a practical option. It was included to provide a hypothetical target for the reject inference methods.

(4) The new accepts sample. This is the extrapolation approach.

(5) The full calibration sample. This is a possible option for reject inference, but one that does not take any account of differences between the new and calibration samples.

(6) Method 7 was used to construct the scorecard with the transformation given by the diagonal matrix **M**.

(7) Method 7 was used to construct the scorecard with the transformation given by (c) from above (where the regression coefficients for the full calibration sample are regressed on the coefficients for the calibration accepts sample).

Table 6.11 shows the proportions of bads at a 70% acceptance rate for the different approaches.

Sample/method	12 characteristics	15 characteristics
1: C(A+R), N(A+R)	20.07	20.36
2: C(A+R), N(A)	20.67	20.59
3: N(A+R)	20.69	20.63
4: N(A)	20.77	20.77
5: C(A+R)	29.42	26.66
6: Method 7 (a)	21.12	21.10
7: Method 7 (c)	21.38	21.25

Table 6.11: A comparison of classification performance for Method 7 and the Bayesian approach to reject inference.

First, the classifier built on the full calibration sample (approach 5) performs very badly relative to the other methods. In fact, it hardly performs better than a random classifier, which would have an expected bad rate of 31.12%. This result suggests that there are fundamental differences between the calibration sample and the new sample. Despite this factor Method *ES* performs better than Method 7.

The second striking feature of the results is that, leaving aside sample/approach (5), the classifiers produced give similar bad rates. This suggests that reject inference is unnecessary for this data set and that the most suitable approach to constructing classifiers is extrapolation.

Unfortunately, Method 7 does not perform well in this comparison given the simplistic assumptions made. It may be that differences between the full and calibration samples were such as to make the method unworkable. If the differences between samples are due to truncation or bias in the calibration sample then it may not be reasonable to expect the relationship between accepts and rejects to be the same in the two samples. This was a fundamental assumption of Method 7. (Alternatively, if the difference between the new and calibration samples is due to a shift in the nature of the applicant population over time, then it may still be reasonable to expect the relationship between the accept and reject samples to remain stable.)

We have used regression coefficients as characterising features of the accept and reject distributions. If one of these coefficients is estimated inaccurately

(perhaps it is close to zero) by the accepts calibration sample then this will give an unstable transformation matrix (M). This causes distortion in the coefficients for the new sample. Further work is needed on the identification of more complex and realistic transformations. One could weight the regression coefficients according to the amount of confidence that one has in them. A comparison of coefficients for different samples would give an impression of the stability of different coefficients.

A second weakness of using regression coefficients as characterising features arises because credit scoring characteristics are highly correlated. This results in unstable coefficients making it very difficult to identify suitable transformations. We could look for alternative characterising features to assess the differences between the calibration and new samples. One way to avoid this problem is to include a window in the transformation that determines the maximum deviation from the original coefficient to the new predicted value. This would reduce the influence of individual characteristics and help to provide a more robust transformation.

In contrast, Method *ES* performs robustly in this experiment. It achieves comparable performance to extrapolation and a classifier built on the full new sample. Because the new sample is relatively larger than the calibration sample, the distorting influence of the calibration sample is reduced.

In conclusion, we have presented a general approach to performing reject inference with a calibration sample. We assessed a simple version of the method and found that it did not compete with extrapolation and Method *ES*. However, this was partly due to the limitations of the data available and the sensitivity of the method to our choice of characterising features.

6.6.4 Method 8 - the mixture-decomposition approach

In Section 6.5.3 we demonstrated that extra information or assumptions are needed if use is to be made of the rejects' characteristic vectors. One such approach that we considered in Section 6.5.4 was to assume distribution forms for $P(\mathbf{x}|g)$ and $P(\mathbf{x}|b)$. The parameters of these distributions are estimated by comparing their mixture distribution with the empirical distribution of $f(\mathbf{x})$. If

the true status of some of the applicants is known then this can be used to improve the estimate, although this is not an essential part of the method.

The weakness of the mixture decomposition approach is that there may not be suitable parameterized distributions for the data. In particular, the assumption of multivariate normality is unrealistic for the credit scoring problem. If we were to use the new accept sample to identify suitable class conditional distributions, the sample truncation might result in the selection of inappropriate distributions for the reject region. Method 8 avoids this problem by using the calibration sample to select suitable families of distributions for $P(\mathbf{x}|g)$ and $P(\mathbf{x}|b)$. The parameters of these distributions are then estimated using the approach described in 6.5.4. This is a much more modest use of the calibration sample than adopted by Methods 6 and 7. This approach has not been tested in practice.

6.6.5 The utilisation of foresight data

So far in this chapter we have considered two sources of additional information to aid the reject inference process: distributional assumptions and a calibration sample including the true status of the rejects. In this section we introduce a novel form of data that is sometimes available in practical situations and can be used in a reject inference procedure. It is called *foresight data* and involves information that becomes available on applicants in a data sample after their application is made (and before the model building procedure is carried out). This information might include notification of a county court judgement for non-payment of another debt or a default with another credit grantor.

The foresight data cannot be incorporated into a scorecard as a standard characteristic because it is never available at the time an application is made (hence the name foresight data). For this reason it is never normally of use to a credit system developer. However, it is a reasonable assumption that the behaviour described by the foresight data, such as defaulting with another company, is highly relevant to the future performance of the applicant. In other words the information may be highly predictive of an applicant's creditworthiness. This motivated the idea that foresight data could be used in

the reject inference procedure to help predict the true status of the rejects and, thus, to allow improved scoring rules to be developed.

Two methods of reject inference which use foresight characteristics are presented.

Method 9

This method is a simple extrapolation approach and does not require a calibration sample. The procedure is to build a scoring instrument on the new accepts sample using foresight and historical data (historical data is the normal data that is available at the application stage). This can then be used to infer the true status of the rejects in the sample. A new scoring instrument is then constructed from the accepts with their true status and the rejects with their estimated good/bad probabilities. The aim is that by using the foresight data we obtain more accurate estimates of the true status for the rejects and thus are able to build improved scorecards. Of course, as already discussed, if we did not use foresight data to infer the status of the rejects then the parameter estimates for the new scorecard would be unchanged from the original ones.

Method 10

This method is a variant of Method 5 considered in 6.6.1 and requires access to a calibration sample. The procedure is to build a scoring instrument on the calibration sample using historical and foresight data and infer good/bad probabilities for the rejects. This model should provide more accurate estimates of the true status of the rejects than the accepts model used in Method 9. As before, a new scorecard is constructed from the accepts with their true status and the rejects with their estimated probabilities.

6.7 A comparison of reject inference methods

This section describes a comparison of some of the reject inference methods considered in this chapter. We applied the techniques to a real data set used for scorecard construction in order to get an idea of their practical feasibility.

The data set used in the comparison consisted of three samples: a calibration sample (Sample A), a "new" sample (Sample B) and a sample from a future time period (Sample C). One of the limitations of the analysis is that Samples A and B were drawn from the same population, which would be unlikely in practice. Sample B is meant to represent the current sample used for scorecard development for which the rejects' true status is unknown. Sample C represents a sample from the population of future applicants. In practice it is the performance of a scorecard on this sample that is most important. The sample sizes for Samples A, B and C are 8000, 24000 and 4000 respectively. A hold-out sample was taken from each of Samples A and B in order to assess the reject inference methods.

The seven methods of analysis are described briefly:

(1) A scorecard was constructed on the calibration sample using linear regression and historical data only.

(2) The scorecard from (1) was used to infer good/bad probabilities for the rejects from Sample B. A new scorecard was then developed from the Sample B historical data using the rejects with their inferred good/bad probabilities and the accepts with their true status (reject inference Method 5 from 6.6.1)

(3) A scorecard was developed on the calibration sample using historical and foresight data. This scorecard could not be used in practice because foresight data would not be available for new applicants, but it does provide an indication of the predictability of the foresight data.

(4) The rejects from Sample B were allocated good/bad probabilities using the scorecard from (3). A new scorecard was then developed from the Sample B historical data using the rejects with their inferred good/bad probabilities and the accepts with their true status (Method 10 from Section 6.6.5).

(5) A scorecard was constructed on the Sample B accepts and this was used to infer good/bad probabilities for the Sample B rejects. A new scorecard was then developed from the Sample B historical data using the rejects with their inferred good/bad probabilities and the accepts with their true status (Method 9 from Section 6.6.5).

(6) A scorecard was developed on the Sample B accepts (extrapolation)

The final scorecards constructed by each method were applied to Sample C and the hold-out samples from Samples A and B with acceptance rates of 50% and 70%. The resulting bad rates are shown in Table 6.12.

	Sample A		Sample B		Sample C	
Method	Accept 50%	Rate 70%	Accept 50%	Rate 70%	Accept 50%	Rate 70%
(1)	12.22	19.22	12.39	19.64	8.39	13.94
(2)	12.78	19.77	12.19	19.71	8.14	14.14
(3)	6.77	16.10	7.73	16.41	3.99	9.27
(4)	12.70	19.64	12.17	19.61	8.36	14.07
(5)	14.04	20.24	13.48	20.24	8.86	14.37
(6)	13.72	20.88	13.06	21.06	8.82	14.49

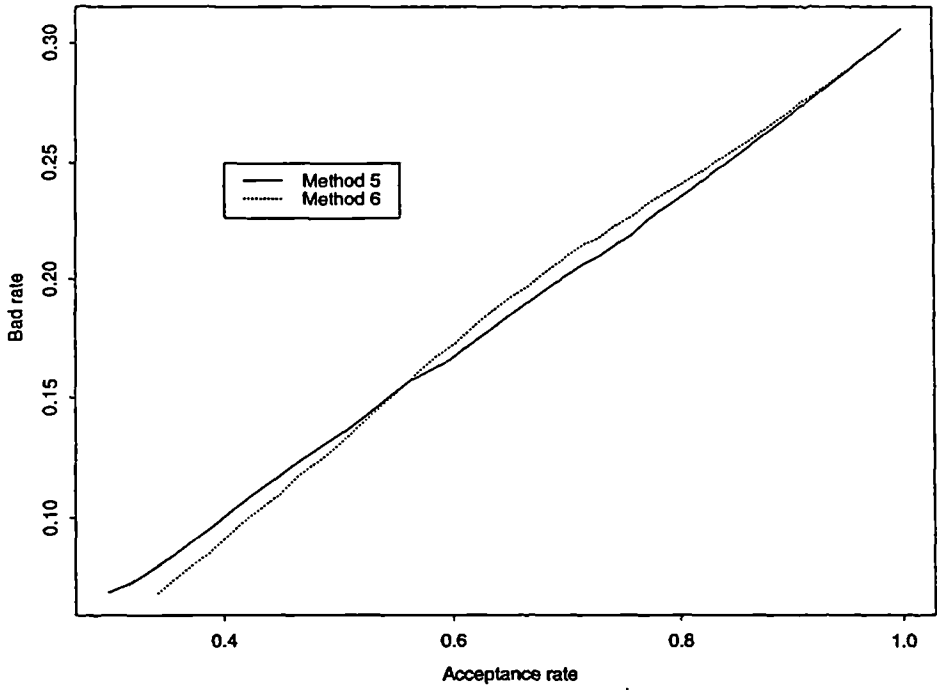
Table 6.12: Bad rates at 50% and 70% acceptance rates for various methods of reject inference.

For reference a scorecard was built on the full Sample B (including the true status for the rejects) and validated on the hold-out part of Sample B. This gave bad rates of 12.11 and 19.30 at acceptance rates of 50% and 70% respectively. Theoretically this should be the best scorecard when tested on Sample B (which by a narrow margin it is). This serves as a convenient baseline against which to compare the other results.

There are several points of interest that arise from the results:

First, when considering Sample B we see that there is little difference between the bad rates from approaches (1), (2) and (4). These methods all have in common the use of a calibration sample. They are all significantly better than

Fig 6.12: Bad rate against proportion accepted for Methods 5 and 6



the results of extrapolation (method (6)). Thus, for this data set, reject inference is necessary and the use of a calibration sample can produce improved results. The limitation of this result is the inherent similarity between the calibration sample and the new sample.

Secondly, we compare the bad rates for methods (5) and (6). For all three samples the extrapolation method does better at the 50% acceptance rate and the foresight data method performs better at a 70% acceptance rate. In order to investigate this phenomenon we plotted bad rate against acceptance rate to give the plot shown in Figure 6.12. This shows that the bad rate curve for Method 6 is below the corresponding Method 5 curve for low acceptance rates and above it for higher acceptance rates. The crossover point occurs when 56% of the sample are accepted.

Looking at the results of using the foresight data in method (3) shows that this data is highly predictive of creditworthiness. The difference between methods (1) and (3) is purely attributable to the foresight data. This difference is far higher than the difference in bad rates between using the full and accept samples (method (6) compared to the results for the full Sample B). This would lead us to expect that the foresight data would allow more accurate prediction of the true good/bad probabilities for the rejects. However, this does not lead to an overall improvement in the predictive accuracy of the model because method (2) and method (4) do not lead to significantly different results. This is a disappointing result.

The explanation is that method (2) is already performing at the same sort of level as a model built on the full sample (our baseline) and so there is no room for an improvement in the bad rate. This result illustrates that new data is more likely to lead to an improvement in bad rate than a new statistical technique.

6.8 Conclusions

Reject inference is the process of inferring the true creditworthiness of rejected applicants for credit. One important reason for needing reject inference is to reduce the sample selection bias that results when a scorecard is constructed using a sample of accepted applicants. We provided an in-depth study into the

nature of the bias and identified several factors which contribute. In particular, we pointed out the necessity of including all the characteristics used to make the original accept/reject decision in the new classifier.

We considered the performance of simple extrapolation methods for reject inference. The success of this approach was seen to depend upon the adopted approach to classifier design: direct methods of estimating $P(g|\mathbf{x})$, such as logistic regression, are much less susceptible to introducing bias than indirect methods such as discriminant analysis. Comparison experiments with more sophisticated reject inference methods showed that extrapolation is surprisingly robust.

We reviewed models proposed in the literature which attempt to use the characteristic vectors for the rejects. A likelihood based approach was used to show that this approach cannot lead to improved parameter estimates unless additional assumptions are made about the class conditional distributions.

A third class of reject inference models was described which makes use of additional information in the form of a calibration sample. These are samples which include the true creditworthiness of the rejects, but may not come from the same population as the current sample. Four methods of using this information were described. In some circumstances these methods were found to reduce the bias of extrapolation methods.

To conclude, we assert that *reliable* reject inference is impossible. In particular, there are three ways in which the performance of an extrapolated model can be improved:

- *Chance.* The new classifier may be better than the old one by luck

- *The use of additional information/assumptions.* Two possible approaches are the assumed distributional forms included in the mixture decomposition method, and information about $P(g|\mathbf{x})$ in the reject region in the form of a calibration sample.

- *Ad hoc adjustment of the accepts classifier.* Expert knowledge of the data may allow the adjustment of parameter estimates in a direction likely to reduce

bias. For example, if a characteristic receives insufficient weighting in a scorecard built on the accepts, it might be possible to reduce bias in the classifier by multiplying the attribute scores by subjective weights.

PART 3

Approaches to classifier design

Chapter 7

A comparison of classification techniques for credit scoring

7.1 Introduction

One aspect of building credit scoring models is the selection of appropriate classification techniques. In this thesis we restrict attention to identifying techniques suitable for discriminating between two populations: the good and bad classes. There are many techniques that have been applied to this problem including discriminant analysis, linear regression, logistic regression and decision trees. More recently other techniques, such as neural networks and genetic algorithms, have been the subject of much interest in the credit industry (see Chapter 4 for a review).

The aim of this chapter is to provide an empirical comparison of a range of classification techniques for credit scoring using a real data set. We divide the discussion into three parts:

(1) In Section 7.2 we focus on a comparison of linear and logistic regression, because of their popularity with credit grantors. This is largely due to their widespread availability in statistical software packages. In Section 7.2.1 we describe why consumer credit data often violates some of the distributional assumptions required by linear regression. In Section 7.2.2 we describe logistic regression, a parametric alternative to linear regression which is more appropriate for binary response data. An empirical comparison considered in Section 7.2.3 provides a surprising result: although the logistic regression classifiers give consistently better performance, the differences are not significant using either of the tests of relative performance discussed in Section 5.3. An explanation for this phenomenon is proposed, based upon the slope of the surface of $P(g|\mathbf{x})$. To assess the nature of any differences we explore how the relative performance changes with acceptance rate and the robustness of the parameter estimates.

(2) In Section 7.3 we extend the comparisons to include other classification techniques that are popular in the statistical literature, but have not been widely applied to the credit scoring problem: projection pursuit regression (see Section 4.6.3) and Poisson regression (see Section 4.7). We also consider decision tree and decision graph classifiers (see Section 4.9). Throughout this chapter we use the performance of linear regression as a baseline for the other methods to beat. Our finding, in line with earlier studies described in Section 3.2, is that linear regression is surprisingly robust to departures from the required distributional assumptions.

(3) In Section 7.4 we assess how the relative performance of the classification methods changes with the definitions of credit default. In particular, this enables us to address the problem of identifying fraudulent applications for credit, an issue of interest to many credit grantors.

In Henley and Hand (1995) we present our conclusions on the selection of classification techniques for credit scoring. We state that sophisticated and sensitive use of any method will lead to performance near the optimum. We also believe that because the relationships between credit characteristics are often monotonic, if not quite linear, the performance of basic statistical methods, such as linear and logistic regression, will be difficult to improve upon.

7.2 A comparison of linear and logistic regression

7.2.1 Linear regression

Linear regression has traditionally been used for building scorecards because of its conceptual simplicity and the widespread availability of statistical software with linear regression routines. The fitted model is given by:

$$E(y_i) = \beta_0 + \beta_1 x_{1i} + \ldots \beta_k x_{ki} \quad (7.1)$$

where $y_i = 0$ if the ith applicant is bad and $y_i = 1$ if the ith applicant is good. Here x_{ji} represents the attribute value of the jth characteristic of the ith applicant. The parameters β_j are estimated by the method of least squares.

The predicted $E(y_i)$ from the model can be interpreted as the predicted probability of the ith applicant being a good risk. To see this we note that if p_i is the probability of the ith applicant being good, then

$$E(y_i) = 1.p_i + 0.(1 - p_i) = p_i \ .$$

It can also be shown that linear regression is equivalent to discriminant analysis when only two groups are used (Leonard (1988), Orgler (1971)). We discussed linear discriminant analysis in Section 4.1.1 and highlighted why it may be inappropriate for building credit scoring models. The related problems of using linear regression to construct credit scoring models are discussed in more detail:

(i) For binary response data a suitable distribution for the ith observation, y_i, is the binomial distribution with parameters $n_i = 1$ and p_i. This means that Var(y_i) = $p_i(1\text{-}p_i)$. Thus, the variance of the observed values is not constant and depends on the unknown p_i. This breaks one of the distributional assumptions of linear regression. The problem can be tackled by using a variance stabilizing transformation or using iteratively weighted least squares to estimate the parameters.

(ii) One of the assumptions of linear regression is that the response variable (and the residuals) are normally distributed. For the reasons mentioned in (i) this is not appropriate for credit scoring data. This assumption is only necessary for tests of significance of individual coefficients and goodness of fit of the model.

(iii) Because the estimated parameters $\hat{\beta}_i$ are unconstrained, the estimated values of p_i are not constrained to lie in the interval [0,1]. Thus, it is inappropriate to use a linear model to predict p_i. This type of probability surface is better approximated by a sigmoid function.

7.2.2 Logistic regression

For the reasons outlined above, linear regression is not a theoretically appropriate method for constructing credit scoring models. There are several types of model that are more suitable for binary response data including logistic regression and probit analysis (see Section 4.7 for a discussion of generalized linear models). They involve transforming the probability scale for the response variable and then fitting a linear model to the transformed values.

In this section we focus on logistic regression because of its widespread use in the credit industry. Furthermore, unlike the probit transformation, logistic regression has a direct interpretation in terms of the logarithm of the odds in favour of being a good. The fitted model is given by:

$$\log(p_i / 1-p_i) = \beta_0 + \beta_1 x_{1i} + \ldots \beta_k x_{ki} \qquad (7.2)$$

where the parameters p_i, x_{ji} and β_j are defined as for linear regression.

The logistic regression model satisfies the criticisms of the linear regression model: first, the response is assumed to be binomially distributed; secondly, the model constrains p_i to lie between 0 and 1 and, thirdly, the curve of logit(p_i) against p_i is sigmoid and symmetric about p_i = 0.5. (However, we note that the curve is essentially linear between for $0.2 < p_i < 0.8$).

Titterington (1992) highlights two difficulties of using logistic regression to build credit scoring models:

(1) There is no explict formula for the ML estimates. As a result it is necessary to use a numerical optimisation algorithm such as the Newton Raphson algorithm or the method of scoring (see Dobson 1983).

(2) There is no standard software routine for building a logistic model in the case where the response (creditworthiness) is assumed to lie on a continuous scale.

Despite these minor considerations one might expect logistic regression to outperform linear regression in the credit scoring domain.

7.2.3 Empirical comparisons of linear and logistic regression

The full sample used in this study is summarised in Table 7.1:

	Number of variables	Number of classes	Number of cases	% of bads in full sample
Full sample	16	2	19186	54.54

Table 7.1: A description of the data set used for assessing the relative performance of linear and logistic regression.

The sixteen characteristics available in the sample were pre-selected from a larger set using the combined set of most predictive characteristics from stepwise linear and logistic regressions.

In this study the bad rates resulting from the linear and logistic regression models are compared using an extension of the hold-out method (see Section 5.2.1.1). This involves repeatedly splitting the full sample into design and hold-out samples (with an 80/20 split), building regression models on the design sample and testing on the hold-out sample. The bad rates are then averaged over hold-out samples. Following Leonard (1988) we use five iterations. The proportion of bads was allowed to vary in the design and hold-out samples, because we are interested in prediction performance on future applicants and the proportion of bads in such a sample will vary.

We split the discussion of the results into three parts:

(1) A comparison of model performance at particular threshold points (corresponding to 30%, 50% and 70% accepts).

(2) A comparison of overall model performance (across all acceptance rates).

(3) An examination of the robustness of the model parameters.

7.2.3.1 Comparison of performance for fixed acceptance rates

In the comparisons that follow the full set of sixteen characteristics were included in both the linear and logistic models. In each case all the characteristics were found to make a significant contribution to the model. Fixed acceptance rates of 30%, 50% and 70% were chosen to represent the type of thresholds that would be used in practice. Figures 7.1 to 7.3 show curves of bad rate averaged over samples against acceptance rate for the linear and logistic classifiers (for particular intervals of acceptance rates about the fixed values). These curves indicate that the two methods give similar relative performance. However, the logistic classifier gives consistently lower bad rates in all three regions of acceptance rate. This consistency gives some grounds for believing there to be a real but small difference in performance.

Table 7.2 displays the bad rates at each of these acceptance rates for the linear and logistic classifiers averaged over samples.

	30% accepted	50% accepted	70% accepted
Linear	18.58	33.39	43.36
Logistic	18.44	33.08	43.30

Table 7.2: Classification results for the linear and logistic regression classifiers averaged over samples.

Although all the differences are small, the logistic classifier does give slightly lower bad rates in each case. The smallest difference in bad rate occurs when a = 70% (the value of interest to the credit grantor in our problem). In order to assess whether the differences in bad rate are due to a genuine improvement in performance we applied the two significance tests of Section 5.3 to each sample. Because the results were consistent for the five samples, we only present the test results for Sample 4.

The classification results for Sample 4 are shown in Table 7.3.

Fig 7.1: Average bad rates in region around a = 30%

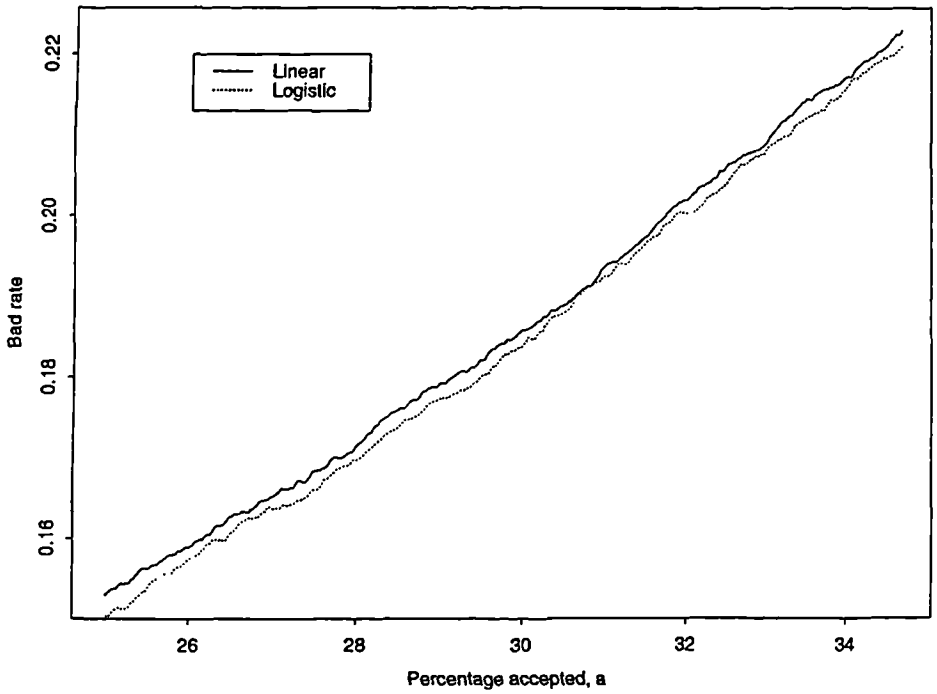


Fig 7.2: Average bad rates in region around a = 50%

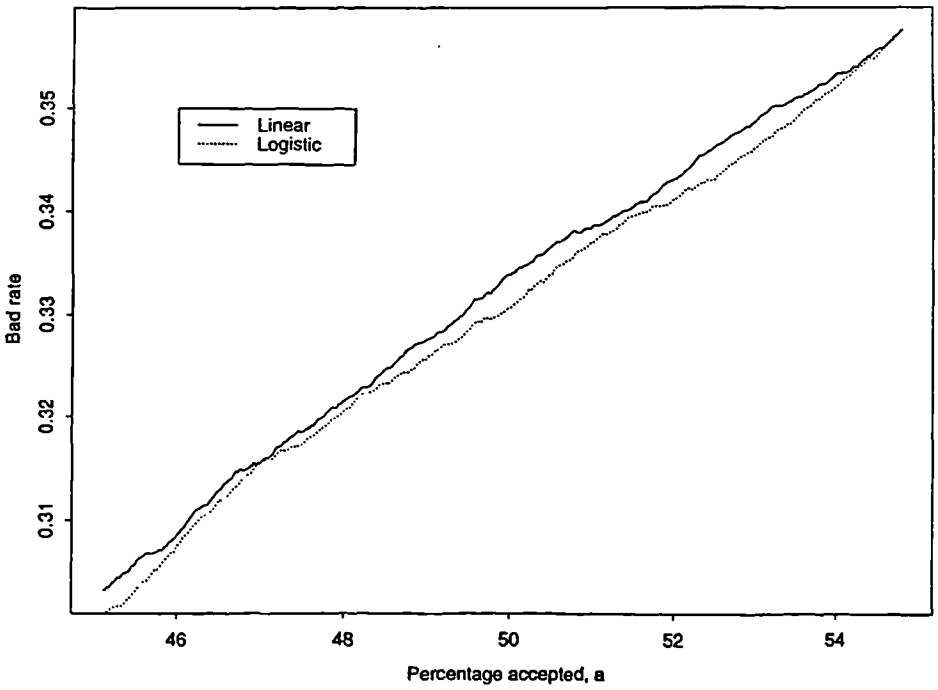
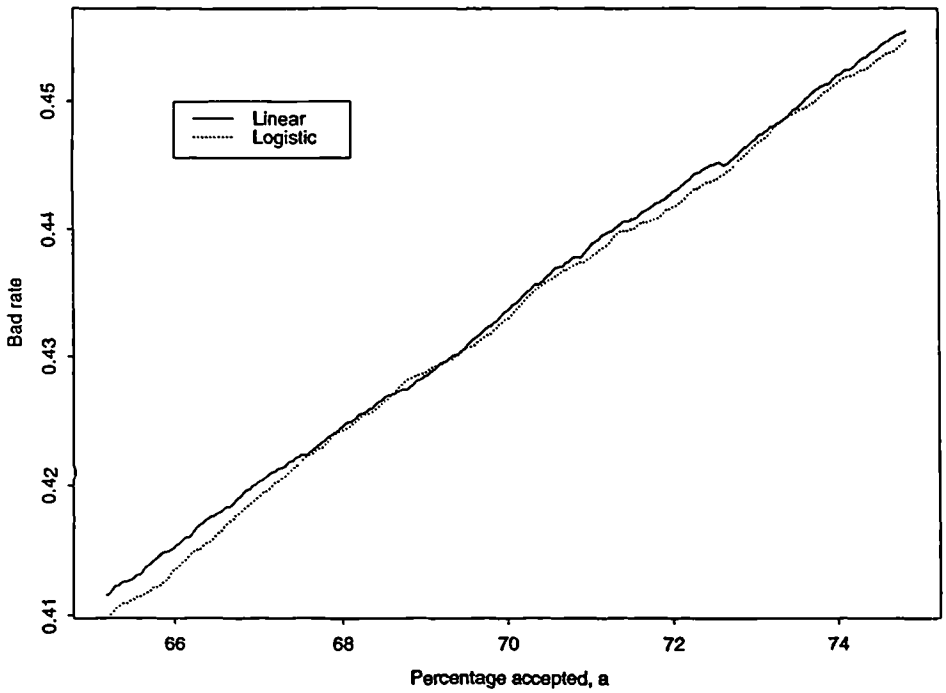


Fig 7.3: Average bad rates in region around a = 70%



	30% accepted	50% accepted	70% accepted
Linear	16.70	32.04	42.16
Logistic	16.27	31.85	42.16

Table 7.3: Classification results for the linear and logistic regression classifiers for Sample 4.

The classification results are equal at the 70% acceptance rate, so we do not need to test for a significant difference in performance. Table 7.4 shows the swapsets for the two classifiers at the 30% and 50% acceptance rates. This table summarises the differences between the accepts under the linear and logistic classifiers.

	30% acceptance		50% acceptance	
	Goods	Bads	Goods	Bads
Linear	9	13	6	15
Logistic	15	7	10	10

Table 7.4: Swapsets for the linear and logistic classifiers for Sample 4.

Table 7.5 shows the resulting *p*-values from (1) the test based upon Fisher's Exact test and (2) the likelihood ratio test.

	30% accepts	50% accepts
Test (1)	0.1292	0.2082
Test (2)	0.6141	0.7984

Table 7.5: Significance test results at 30% and 50% acceptance rates.

Using either test we come to the same conclusion: for this sample there is not a statistically significant difference between the performance of the linear and logistic regression classifiers. (The big difference between the *p*-values for the two tests results from the different hypotheses tested in each case. Both tests can be useful in assessing relative performance for the reasons discussed in the conclusion to Chapter 5.)

It was also found that there was not a significant difference between the linear and logistic classifiers for the other four samples considered individually, at any of the fixed threshold points. These results are surprising because, as argued in Section 7.2.2, logistic regression is theoretically more appropriate than linear regression for consumer credit data.

Although the differences in performance for each sample are not significant, the logistic classifier does give consistently better average bad rates (see Table 7.2). Table 7.4 shows that if there is a real difference in performance, then logistic regression is able to replace about five misclassified bads in the accept region with goods. Although this is a relatively small improvement, one bad applicant can cost the credit grantor a considerable sum of money. Therefore, if faced with a choice between using linear and logistic regression to build classifiers, it would seem more sensible to opt for the logistic classifier.

7.2.3.2 Comparison of overall performance

Figure 7.4 shows curves of bad rate against acceptance rate for the linear and logistic regression models for Sample 1. The best and worst bounds on performance are added. This indicates that the two models give very similar relative performance over the entire range of acceptance rates. The graph shows that the relatively high bad rates for high acceptance rates are close to the best that can be achieved. Thus, the models are providing (at least) reasonable discrimination between goods and bads.

Figure 7.5 shows corresponding curves for Sample 3. The only noticeable difference between the curves for Samples 1 and 3 is the sharp peak for very low acceptance rates for Sample 3. Although the bad rate jumps to 0.2, the peak is caused by the acceptance of just one bad applicant. The large jump in bad rate results because of the low acceptance rate.

Figure 7.6 shows the bad rate curves for linear and logistic regression averaged over samples. This shows that the effect of the blip in the low acceptance region for Sample 3 has been diminished by the effect of the other samples. The average curves for linear and logistic regression are very similar across the entire range of acceptance rates indicating similar overall performance. To

Fig 7.4: Linear and logistic bad rate curves for Sample 1

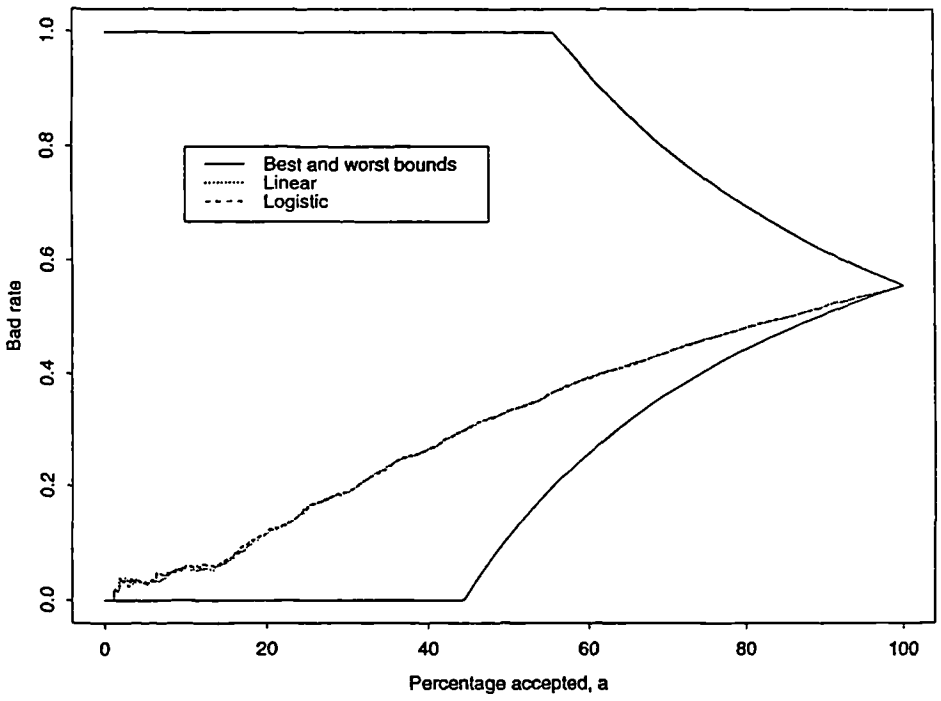


Fig 7.5: Linear and logistic bad rate curves for Sample 3

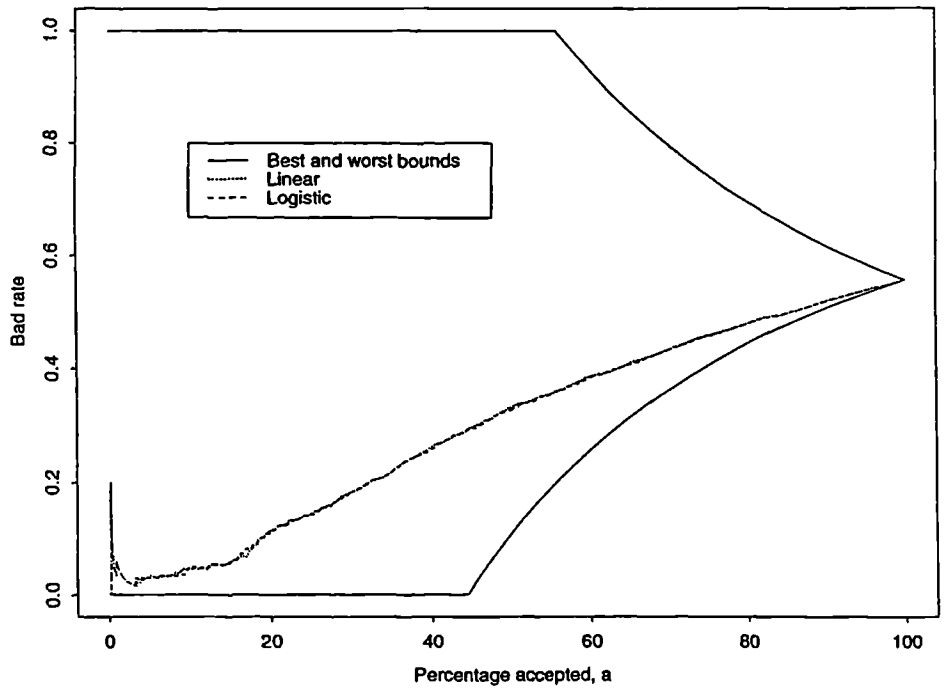


Fig 7.6: Linear and logistic bad rate curves averaged over samples

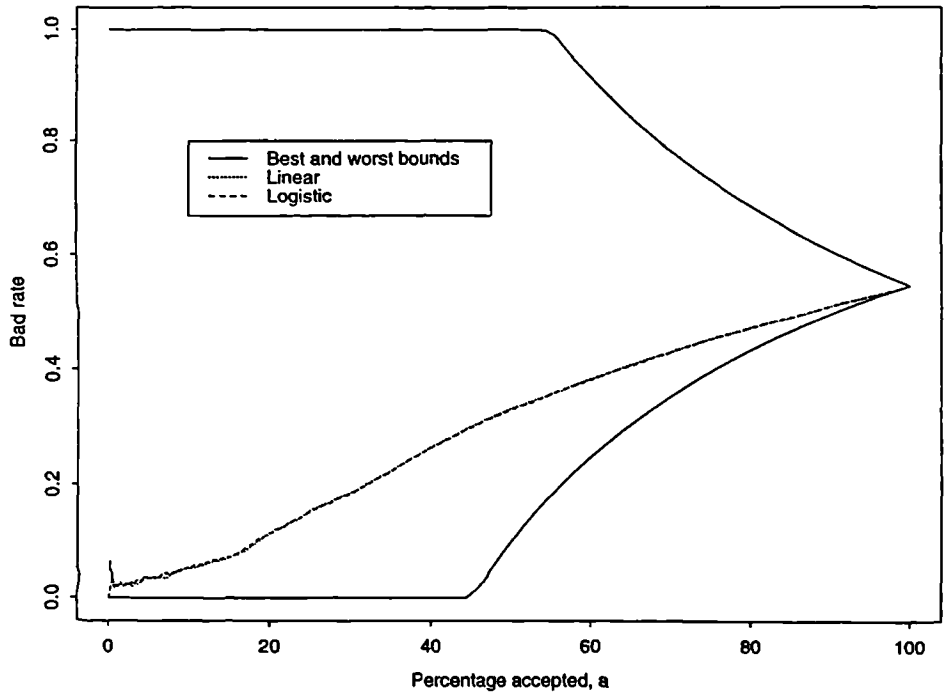


Fig 7.7: Confidence limits for linear regression

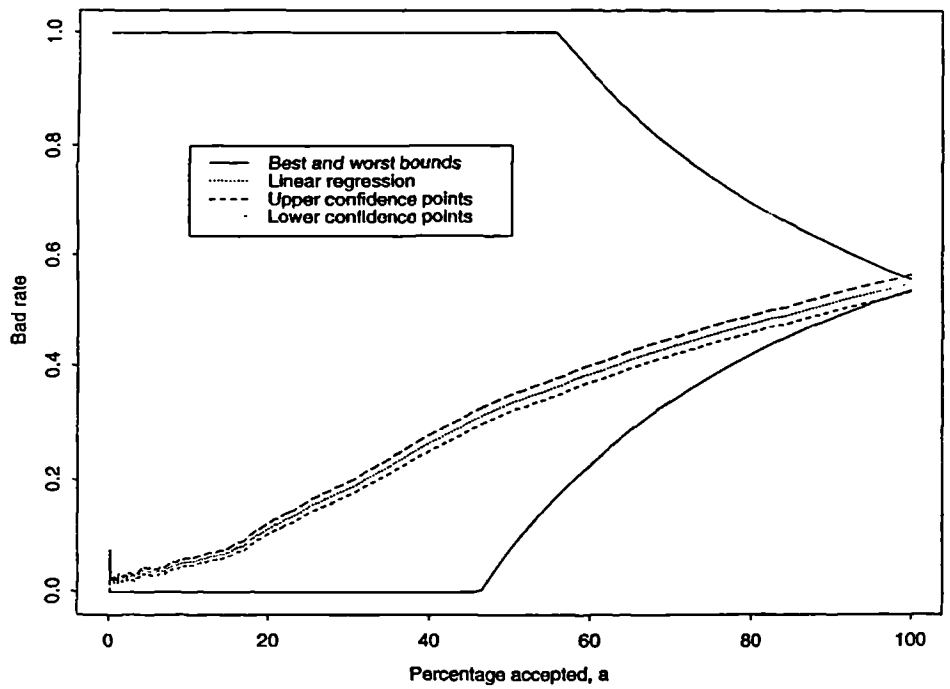


Fig 7.8: Confidence limits for logistic regression

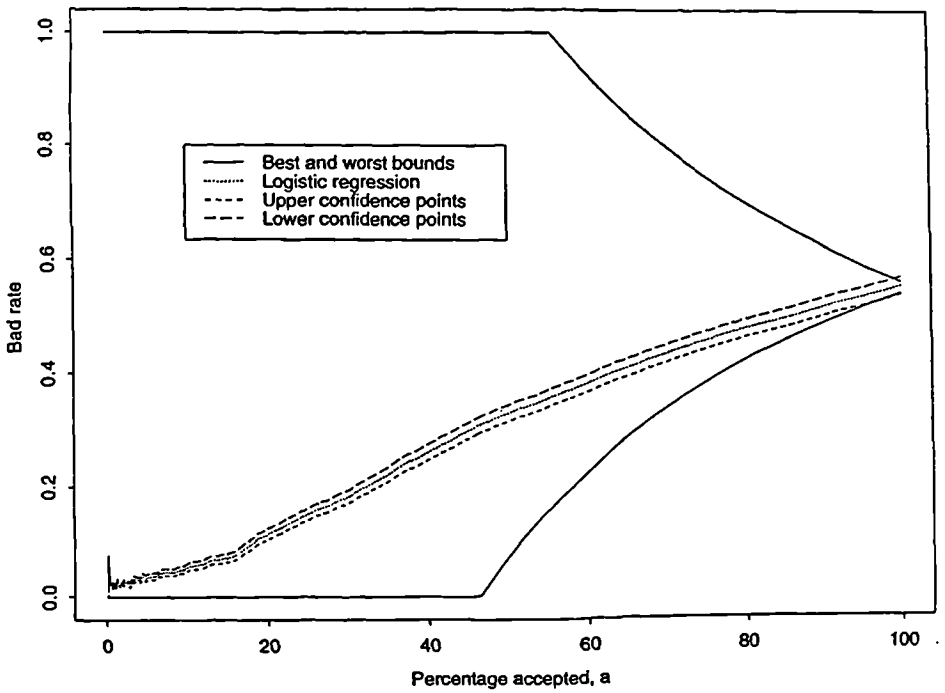
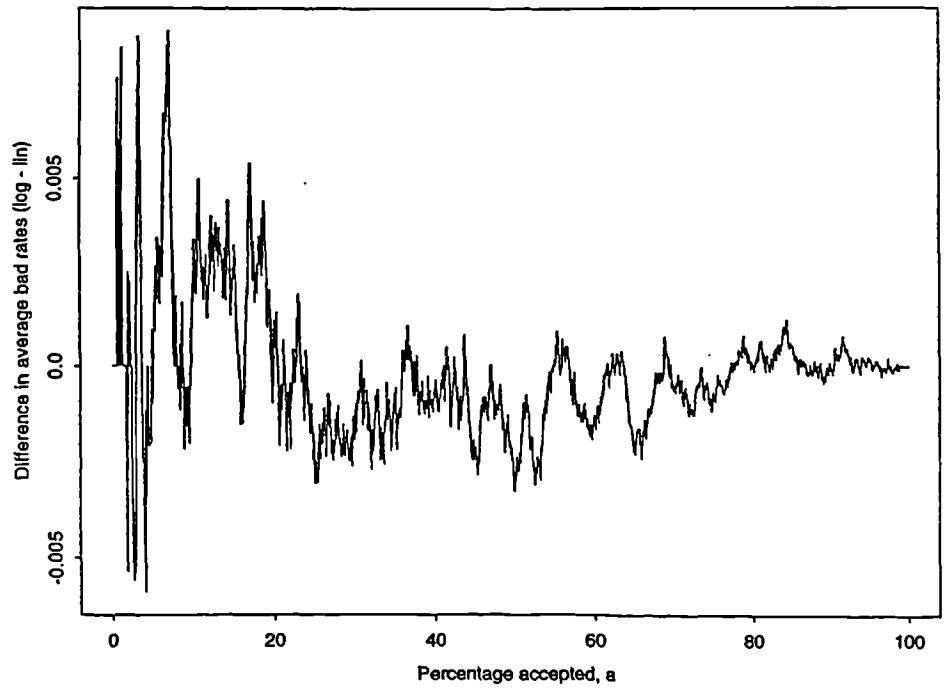


Fig 7.9: Difference in average bad rate against acceptance rate



assess the consistency of these curves, Figures 7.7 and 7.8 show the averaged bad rate curves for linear and logistic regression respectively with upper and lower 95% confidence points. The confidence limits are almost identical for the two curves. The difference between the linear and logistic curves is negligible compared to the width of the confidence intervals.

To assess the nature of any differences in performance, Figure 7.9 shows the difference in average bad rate (logistic minus linear) against acceptance rate. There are two interesting features of these results: first, the graph clearly shows how the difference in performance between the two classifiers diminishes as acceptance rate increases. Secondly, for acceptance rates above about 20% the logistic classifier gives slightly better performance on average than the linear classifier. For acceptance rates below 20% the differences in performance fluctuate erratically from positive to negative. The sponsors of this project place particular emphasis on an acceptance rate of 70%. At this acceptance rate, as discussed above, there is little difference between the two classifiers, but the logistic classifier might be preferred.

Another approach to assessing the overall performance of a credit scoring model, and one that is commonly used in the credit industry, is the Lorentz diagram and the Gini coefficient. Figure 7.10 shows the Lorentz curve for the linear regression classifier on Sample 2. The logistic curve was found to be indistinguishable from the linear curve. (This was consistent across the five samples). The Gini coefficients are given by 0.58399 for the linear classifier and 0.58407 for the logistic classifier. The similarity of these two values gives further evidence of the lack of substantial difference in relative performance.

7.2.3.3 Robustness of the parameter estimates

Although the results of the last two sections showed that there was not a statistically significant difference between the performance of the linear and logistic classifiers, Figure 7.9 gave some indication that the logistic classifier performs slightly better for acceptance rates above 20%. In this section we further explore sources of differences between the two classifiers in order to aid a decision as to which technique to adopt. We consider two possible criteria for choosing between techniques:

Fig 7.10: Lorentz curve for linear regression on Sample 2

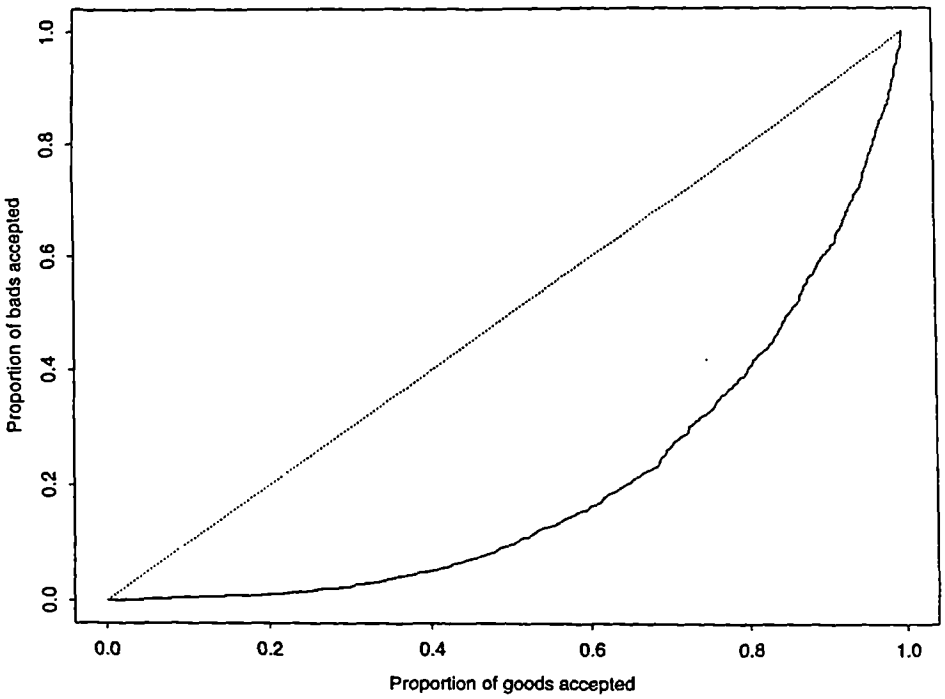
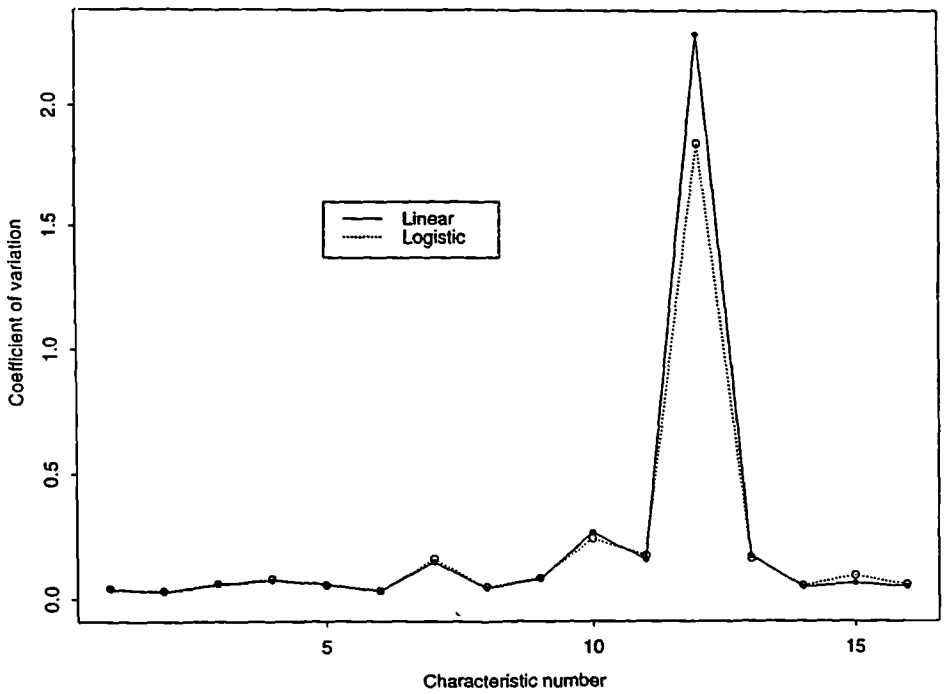


Fig 7.11: The coefficient of variation for different characteristics



(1) *The robustness of the model to sampling variation.*

If a technique provides stable model coefficients then the credit grantor will have more confidence in applying the model to similar populations. One way of assessing the model stability is to consider the coefficient of variation for each characteristic. If the mean and standard deviation of the model coefficients for the five samples used above are given by $\bar{\mathbf{x}}_i$ and $\mathbf{s}_i$, where i corresponds to the characteristic number, then the coefficient of variation for characteristic i is given by:

$$C_i = \frac{\mathbf{s}_i}{\bar{\mathbf{x}}_i}.$$

Figure 7.11 shows the coefficients of variation for the sixteen characteristics included in the linear and logistic models. Both models have relatively low variation for fifteen of the characteristics. The twelfth characteristic has a much higher coefficient of variation because it is the least important characteristic in the model and takes a negative value for some of the samples. The plot gives further evidence of the strong similarity of the two models.

Another approach to assessing the stability of the model coefficients is to consider boxplots for individual characteristics. Figure 7.12 shows such plots for a selection of four characteristics. Again the similarity of the sample distributions for the linear and logistic classifiers is apparent.

(2) *The relative positioning of the model coefficients.*

A credit grantor is likely to prefer a model which gives emphasis to characteristics which experience has shown to be highly correlated with creditworthiness and which appear sensible. As a result we looked for differences in the relative importance of characteristics in the linear and logistic models. Table 7.6 shows the ratio of each model coefficient (averaged over samples) to the largest average coefficient for the linear and logistic models. In Figure 7.13 the relative ratios are plotted against the characteristic number. As expected the curves are very similar indicating that the relationships between individual regression coefficients are indeed similar for the two models. This results in similar classification performance under the two models. In the next section we propose an explanation for why the fitted models have this similar form.

Fig 7.12: Boxplots for characteristics 1, 3, 9 and 16

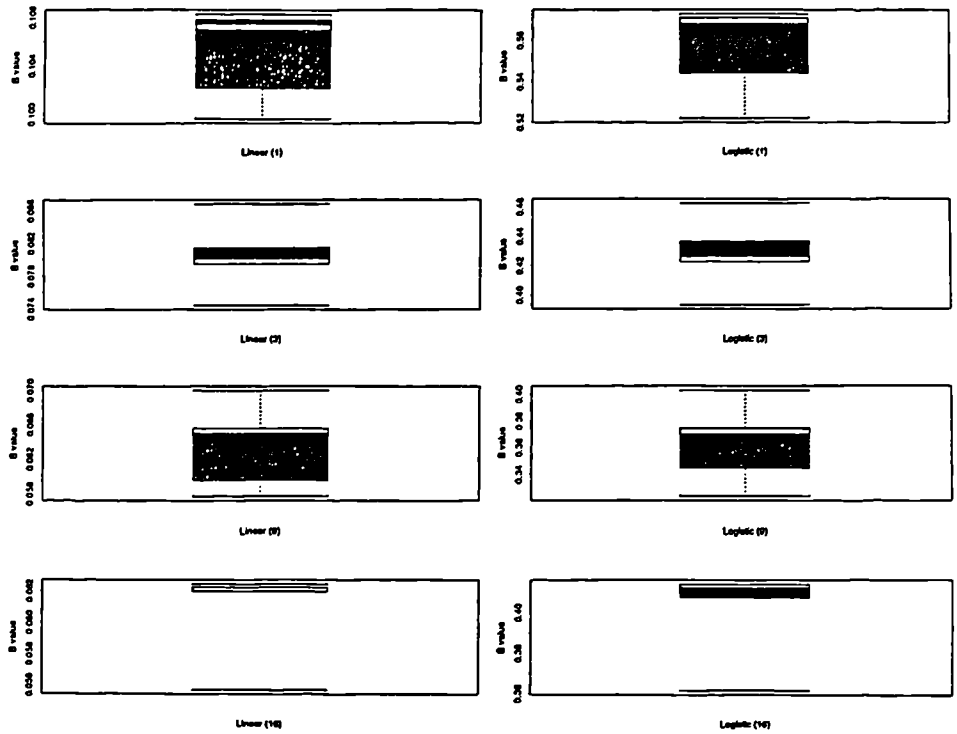
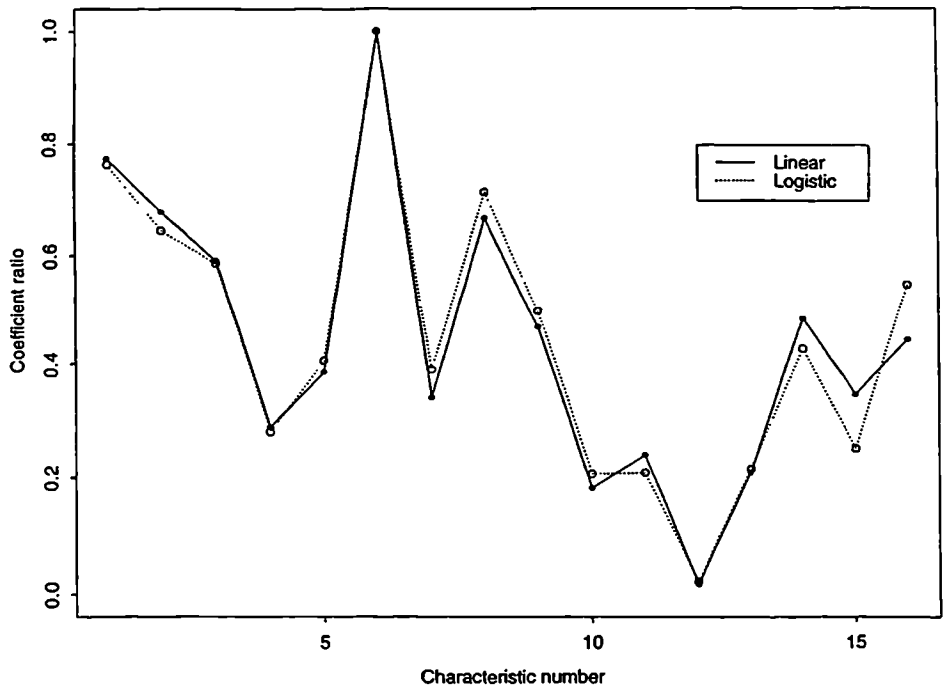


Fig 7.13: Ratio of regression coefficients to largest coefficient



Char. no	Linear	Logistic	Char. no	Linear	Logistic
1	0.774	0.764	9	0.469	0.499
2	0.679	0.646	10	0.182	0.207
3	0.593	0.588	11	0.241	0.209
4	0.289	0.280	12	0.017	0.021
5	0.387	0.407	13	0.211	0.217
6	1.000	1.000	14	0.487	0.431
7	0.342	0.391	15	0.349	0.253
8	0.669	0.716	16	0.449	0.552

Table 7.6: Ratio of model coefficients to the largest coefficient for linear and logistic regression.

7.2.3.4 An explanation for the similar relative performance

In Section 8.5.4.2 we justify with the aid of suitable plots that the slope of the true $P(g|\mathbf{x})$ is relatively shallow across the characteristic space. This means that the true curve of $P(g|\mathbf{x})$ is between 0.2 and 0.8 for a relatively high proportion of the population. As mentioned earlier, the logistic curve is approximately linear in this region. This means that the predicted $P(g|\mathbf{x})$ from the linear and logistic models will be similar for a large proportion of the sample. (It is only when there are substantial numbers of applicants with $P(g|\mathbf{x})$ close to 0 or 1 that the fitted models will differ substantially.) Thus, the two models will give similar classification accuracy. Any small but real improvement in performance can be attributed to the increased accuracy of the logistic model in the regions of extreme $P(g|\mathbf{x})$.

7.3 Comparisons with other classification techniques

The aim of this chapter is to identify techniques that give better classification performance than linear regression in empirical comparisons. The results of the previous sections indicate that despite the theoretical advantages of logistic regression over linear regression, this is not reflected in a substantial difference in classification performance for our data set. More generally we believe that varying the classification technique will not lead to big changes in relative

performance of credit scoring models. In this section, to test this hypothesis out, we assess the relative performance of a range of other classification methods. Poisson regression (see Section 4.7) and projection pursuit regression (see Section 4.6.3) were included to represent other suitable parametric and non-parametric regression techniques. Neither of these techniques has received previous treatment in the credit scoring literature. We also include comparisons with decision trees and decision graphs (see Section 4.9). Decision trees were included because of their popularity in the credit scoring field and decision graphs were included to represent a recent development in classification methodology. In what follows we take the performance of linear regression as a baseline for the other methods to beat.

In all the credit scoring models constructed in this comparison we use the same sixteen characteristics used in the previous section. Although this may give a small advantage to the linear and logistic classifiers, it reduces the number of variable factors in our experiment. A fixed acceptance rate of 70% was adopted because of its interest to the credit grantor in our problem. In the previous comparisons the main effect of acceptance rate on relative performance was on the strength of the result, rather than the result itself.

Table 7.7 shows the bad rates at a 70% acceptance rate for the range of techniques described above for each sample. Because the decision graph method includes a metric which indicates whether a graph or tree is more appropriate to the data, we present the results of these two methods together.

Sample	Linear	Logistic	Projection pursuit	Poisson	Decision graphs/trees
1	43.87	43.92	43.79	43.91	44.09
2	44.03	43.94	43.89	44.11	44.50
3	43.54	43.45	43.47	43.54	44.19
4	42.16	42.16	42.16	42.14	42.60
5	43.20	43.01	43.03	43.09	43.46

Table 7.7: Bad rates at a 70% acceptance rate for a range of classification techniques.

As in Section 7.2 none of the differences in bad rate between two classifiers are significant using the two significance tests from Chapter 5. Despite this, the decision tree/graph results are consistently worse than the results for other classification techniques. One possible explanation for this difference is the inclusion of a decision tree as a characteristic (because this may indirectly give the advantages of the decision tree approach to the other classification methods).

Table 7.8 shows the bad rates averaged over samples for the range of classification techniques considered.

Method	Bad rate
Linear regression	43.36
Logistic regression	43.30
Projection pursuit	43.27
Poisson regression	43.36
Decision graphs/trees	43.77

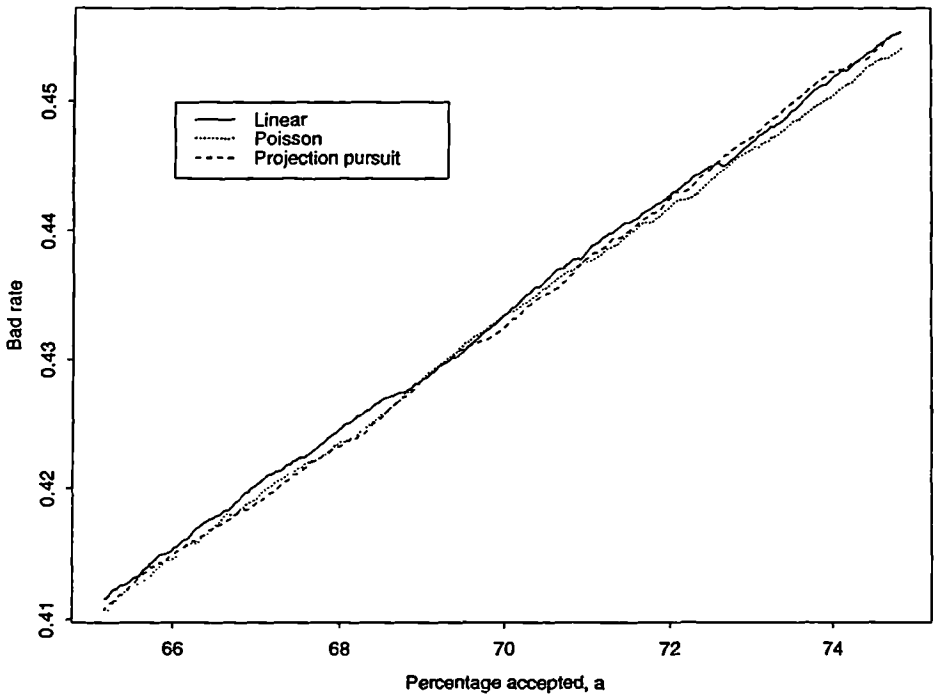
Table 7.8: Bad rates averaged over samples for different classification techniques.

The table shows the expected performance of the different classification techniques when applied to the population from which our sample is drawn. Projection pursuit regression and logistic regression give slightly better average performance than linear regression, although the differences in performance are not statistically significant for samples tested individually. Even though this difference is small, if real it could result in large savings for the credit grantor.

In order to assess further the nature of any differences in performance between classification techniques, Figure 7.14 shows the averaged bad rate curves for linear regression, Poisson regression and projection pursuit regression for a range of acceptance percentages centred on 70%. This shows that relative performance fluctuates with acceptance rate in an apparently random manner.

We conclude that there is no discernible real difference in classification performance for the range of parametric and non-parametric techniques considered. This supports our assertion at the beginning of this section that

Fig 7.14: Average bad rates in region around a = 70%



varying the classification technique does not lead to large changes in performance. However, if a choice has to be made about the appropriate classification technique to adopt for building credit scoring models, linear and logistic regression appear to be the most suitable of those considered because of their inherent simplicity.

7.4 The influence of alternative definitions of credit default on classifier performance: fraud scoring

7.4.1 An introduction to fraud scoring

The area of fraud is becoming of increasing interest to credit grantors. Leonard (1993/4) describes an expert system model designed to alert banks and other financial institutions to the fraudulent use of credit cards. This type of fraud can arise in two ways: *counterfeit fraud* occurs when an unauthorized duplicate of the credit card is in circulation and *non-receipt fraud* occurs when a card is intercepted and used by a third party. The proposed model attempts to detect these types of fraudulent activity by identifying deviations from normal spending patterns. This allows the credit grantor to react by contacting the customer and, if necessary, blocking any futher use of the account. Leonard describes an experiment to compare the expert system with a naive model which classifies all customers as either good accounts or fraudulent accounts. The results show that, taking into account different costs of misclassification, the fraud alert model can lead to substantial savings (estimated at $245,183 per month).

In this section we consider a different type of fraud that is more specific to the mail order industry. This occurs when a customer spends up to the allocated credit limit and then makes no repayments. This can be viewed as fraud because the customer exhibits no willingness to repay. For operational reasons we simplify the definition slightly: a fraud is defined to be a customer who makes no repayments but does not necessarily spend up to the credit limit. In Section 7.4.2 we describe the construction of classifiers to discriminate between frauds and non-frauds and make comparisons with standard good/bad risk classifiers.

The approach we take can be seen as a way of varying the definition of credit default. Our definition of fraud can be seen as a laxer definition of a bad risk. A similar approach is taken by Crook et al. (1992). They make a comparison of the ranking of the predictor variables under the standard definition of credit default used in this thesis (see Section 2.1) with a more stringent definition. The stringent definition of a bad risk was an applicant who subsequently missed at least one repayment.

The authors constructed credit scoring models using discriminant analysis on a sample of 1001 applicants for a credit card with 23 characteristics available. A few characteristics were deleted before carrying out the model construction in order to reduce the correlations between the predictor variables. This was especially important in this study because highly correlated predictor variables lead to unstable model coefficients, thus making it difficult to assess the relative contribution and true ranking of each characteristic. The final models obtained using the two definitions of credit default were then assessed on a 20% hold-out sample. The proportions correctly classified by the discriminant models were compared with the proportional chance criterion:

$$C_{prop} = p^2 + (1-p)^2$$

where p is the proportion of applicants in one of the classes (e.g. goods).

The results show that for both definitions of credit default the discriminant models give better classification accuracy than would be expected by chance. The rankings of the predictor variables were examined for the two models and useful differences were identified. We briefly discuss three points arising from this paper of relevance to our study of fraud:

(1) The analysis was limited by omitting applicants who did not use their credit card in the observation period. This group corresponds to the "other" class in our problem. As discussed below, we choose to include the "others" in the model construction as goods.

(2) The assessment of classification performance was limited to comparing the discriminant models with the chance criterion C_{prop}. In Section 7.4.2 we include a comparison of different regression classifiers to assess whether the appropriate classification method varies with the definition of creditworthiness.

(3) It is not usually sensible to make a direct comparison of classification models constructed using different definitions of credit default. This is because there is no objective definition of creditworthiness that can be used in the comparison. Thus, it is difficult to assess whether a new definition of credit default allows a better representation of credit behaviour. (The paper by Crook et al. (1992) was not aiming to do this.) In the fraud context, if one is happy with the chosen definition of fraud, then one can use this as the "objective" response definition to enable suitable comparisons.

7.4.2 Fraud classifiers

The analysis described in this section is divided into two parts: first, we attempted to construct linear regression classifiers which could identify frauds more effectively than standard risk classifiers; secondly, we considered whether using an alternative definition of credit default (the fraud definitions) led to any change in the relative performance of the classification techniques discussed in Sections 7.2 and 7.3.

The following definition of fraud was used in our comparisons: applicants who made no repayments were classified as fraudulent (the "bads"). All other applicants, including applicants who were normally placed in the "other" classes, were classified as non-fraudulent (the "goods"). The broad definition of a non-fraud was chosen because all the applicants in this class showed some willingness to repay by making at least one repayment.

7.4.2.1 A comparison of fraud and standard risk classifiers

The samples used in this study are summarised in Table 7.9:

Sample	Number of variables	Number of classes	Number of cases
Analysis	81	2	11216
Validation	81	2	2784

Table 7.9: A description of the data sets used for constructing and assessing

fraud classifiers.

Fraud and risk classifiers were constructed using stepwise linear regression on the analysis sample with the full set of 81 characteristics and the appropriate definition of credit default. In each case twenty characteristics were included in the final classifier.

The two classifiers were compared using the validation sample with the fraud definitions. Model performance was assessed by considering the fraud rate amongst accepted applicants for different acceptance rates. Figure 7.15 shows the overall curve of fraud rate against acceptance rate. This shows how the classifier achieving the best performance varies with acceptance rate: for acceptance rates between 30% and 50% the risk classifier is more successful at rejecting frauds, whereas the fraud classifier is more successful at rejecting frauds for acceptance rates between 60% and 80%. It is surprising that a classifier constructed using inappropriate response definitions (the risk classifier) outperforms a classifier constructed using appropriate response definitions (the fraud classifier). Further testing is needed with other samples to test whether this result is due to sampling variation. The results do give some evidence that omitting the "other" classes from the analysis sample (as is done in the construction of risk classifiers) does not lead to a big change in the resulting performance of the classifier.

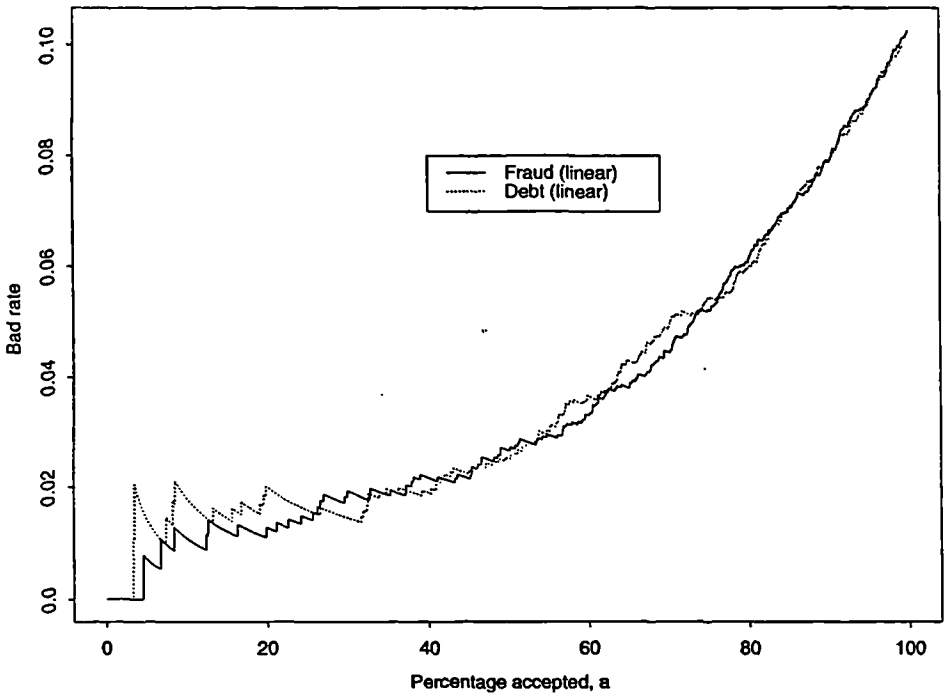
Because of the criterion of the credit grantor we focused attention on an acceptance rate of 70%. The corresponding fraud rates amongst the accepts were 4.93% for the risk classifier and 4.46% for the fraud classifier. The two significance tests of Section 5.3 were used to assess whether these rates are significantly different. Table 7.10 shows the swapsets for the two classifiers.

	Goods	Bads
Risk	167	29
Fraud	176	20

Table 7.10: Swapsets for the risk and fraud classifiers.

The resulting p-values from (1) the test based upon Fisher's Exact test and (2) the likelihood ratio test are 0.11 and 0.65 respectively. We conclude that there

Fig 7.15: Comparison of fraud and debt definitions of credit default



is not a significant difference between the performance of the two classifiers at a 70% acceptance rate. If this result is robust for other samples then there is no advantage in building specific fraud classifiers to identify fraudulent applications for credit in our problem.

7.4.2.2 A comparison of different classification techniques for fraud scoring

Further comparisons were carried out to assess whether the relative performance of the techniques considered in Section 7.3 is sensitive to the definitions of credit default adopted. To achieve this fraud classifiers were constructed using the definition of fraud defined above. Table 7.11 shows the resulting fraud rates amongst the accepts in the validation sample at an acceptance rate of 70%.

Method	Fraud rate
Linear regression	4.46
Logistic regression	4.11
Projection pursuit	4.31
Poisson regression	4.62

Table 7.11: Bad rates averaged over samples for different classification techniques.

None of the differences in fraud rate were found to be significant using either test from Section 5.3 with a significance level of 5%. This result is consistent with the corresponding conclusions of Section 7.3.

7.5 Conclusions

In this chapter we have constructed credit scoring models using a range of parametric and non-parametric classification techniques: linear regression, logistic regession, Poisson regression, projection pursuit regression, decision trees and decision graphs. The results have shown that, given our data set:

- Linear regression is surprisingly robust to departures from the required distributional assumptions.

- There is not a statistically significant difference between the performance of the range of techniques considered. If this insensitivity of classifier performance to the technique used to build it can be confirmed for other data sets then this has implications for further research into credit scoring methodology. It indicates that research time may be more profitably spent on other aspects of the credit granting process such as identifying new sources of data.

- The results are insensitive to whether the fraud or risk definitions of credit default are used. However, future work should assess the influence of more radical innovations on classifier performance, such as use of continuous definitions of credit default.

Chapter 8

Application of the *k*-Nearest Neighbour method to credit scoring

8.1 Introduction

Traditional credit scoring methodology has focused on using techniques such as discriminant analysis and linear and logistic regression to discriminate between good and bad applicants for credit. As has been discussed in earlier chapters, these techniques can provide good discrimination between the good/bad classes. However, they all suffer from the disadvantage of assuming an overall parametric model form (usually linear) for $P(g|\mathbf{x})$. Two aspects of consumer credit data give reason to question the appropriateness of these assumptions: the high correlations between variables (resulting in complex non-linear interactions) and the categorical nature of the data. This has prompted researchers to consider alternative classification procedures for assessing consumer creditworthiness, which do not make restrictive parametric assumptions. Many non-parametric classification techniques have been developed over the last thirty years (e.g. *k*-Nearest Neighbour, kernel methods, neural networks, decision trees) and the emergence of new computer technology has allowed their application to large data sets. In this chapter we consider the application of the *k*-Nearest Neighbour method (*k*-NN) method to credit scoring.

The *k*-NN method is a standard non-parametric technique used for probability density function estimation and classification originally proposed by Fix and Hodges (1952) and Cover and Hart (1967). It was introduced in Section 4.7 as a specific type of variable kernel density estimator. It was chosen as a suitable method for applying to consumer credit data for several reasons:- the non-parametric nature of the method may allow subtle aspects of the data to be modelled that cannot be identified by a parametric method such as logistic regression; the *k*-NN method has been found to perform better than other non-parametric techniques such as kernel methods when the data is multi-

dimensional (Terrell and Scott (1992)); the k-NN method is a fairly intuitive procedure and as such could be easily explained to business managers who would need to approve its implementation. A simple NN classifier was applied to credit scoring data in Fogarty and Ireson (1993) in a general comparison of techniques. The results were promising despite the simplicity of the NN classifier used.

The k-NN method involves estimating the good/bad probabilities for an applicant to be classified by the proportions good and bad amongst the k "most similar" points in a training sample. The similarity of points is assessed using a suitable distance metric. An important part of our analysis will be the selection of suitable distance metrics for the k-NN method. We begin by discussing the literature on the k-NN method in Section 8.2, with particular emphasis on distance metrics. An adjusted version of the Euclidean metric which incorporates knowledge of underlying equi-probability contours for class membership is proposed in Section 8.3.1. These contours are estimated using regression weights, calculated from the data, with the intention of incorporating into the metric knowledge of class separation in the sample. A general transformation of the data which has the same property is defined in Section 8.3.2. This allows one to transform the data before using any standard metric, thus providing a general framework for selecting a data dependent metric. We consider a range of possible metrics including the city-block and the general Minkowski distance.

The aim of this chapter is to provide a practical classification model that can improve upon traditional credit scoring techniques and so in Sections 8.4 and 8.5 we describe the application of our k-NN methodology to a data set supplied by a large mail order company. In Section 8.4 we consider issues relating to the implementation of the method. In Section 8.5 we describe a simplified investigation of the performance of the k-NN method, designed to assess whether it is likely to be competitive with other methods. Our conclusions are

- the k-NN method with adjusted Euclidean metrics shows the potential to give better performance than a range of traditional credit scoring techniques when the two k-NN parameters (k and a distance parameter, D) are estimated using a simplified approach (which cannot be used in practice).

- the k-NN method with the adjusted Euclidean metric gives better performance than the method with the standard Euclidean metric and similar performance to the "optimal" versions of the general Minkowski distance.

- plots of bad rate against k showed that the k-NN method is fairly insensitive to the choice of parameters k and D and, in particular, the curves have surprisingly flat valleys.

- other interesting features of the results include the high optimal k.

- although the bias of the estimates of $P(g|\mathbf{x})$ from the k-NN method increase as k increases, the consequent deterioration in performance is slow due to the shallow slope of $P(g|\mathbf{x})$.

This is followed by a more in-depth study in Section 8.6, which includes a practical procedure for estimating the parameters k and D, as well as a more robust evaluation of performance using several design/test resamples from the data (see also Henley and Hand (1994) and Henley and Hand (1995)). We include comparisons with a range of other classification techniques including linear regression, logistic regression and decision trees. We conclude that

- the k-NN method with adjusted Euclidean metrics gives better performance than the range of other classification techniques in practical comparisons. However, the differences are not statistically significant.

- an appropriate method for selecting the k-NN parameters is to select the values of k and D which minimise the smoothed curves of bad rate against k.

In Section 8.7 we discuss a second comparison study using a test sample from a future population. The idea was to evaluate the robustness of the k-NN method to changes in the nature of the population over time. It is found that

- the k-NN method is less robust to population changes than linear regression.

- a hybrid of the k-NN and linear regression classifiers can lead to improved performance over the individual classifiers.

8.2 Review of k-NN methodology

8.2.1 Description of the k-NN classifier

The theory of the k-NN method is explained in many text books including for example Hand (1981). We will briefly outline the essential details of the method. Using the notation of Chapter 2, the aim of the method is to obtain estimates of $P(i|\mathbf{x})$ and assign the point $\mathbf{x}$ to the class i for which this estimate is largest. One starts by finding the k "most similar" points to the point $\mathbf{x}$ to be classified (what we mean by similarity is described below). Here k is an integer parameter that is chosen by the model developer. An estimate of $P(i|\mathbf{x})$ is given by k_i/k, where k_i is the number of applicants from class i that are contained in the nearest k points. The classification rule is as follows: classify $\mathbf{x}$ as belonging to class j if $k_j = \max_m(k_m)$. Costs could be included in the classification rule if we knew that the implications of the different types of misclassification were different (and we knew the costs).

In order to measure the similarity of two points $\mathbf{x}$ and $\mathbf{y}$ we use a suitable distance measure or metric, denoted by $d(\mathbf{x},\mathbf{y})$. To find the k most similar points to $\mathbf{x}$ we calculate the value of the distance metric $d(\mathbf{x},\mathbf{y})$ for all the applicants $\mathbf{y}$ in the training set and select the k points with the lowest value.

We use a slight variation on this classification rule, due to our criterion for measuring performance. Instead of assigning a point with characteristic vector $\mathbf{x}_i$ to the class to which the majority of its k-nearest neighbours belong, we choose to accept (i.e. classify as being good) a certain proportion of the sample on the basis of ranking the estimated $P(g|\mathbf{x})$. This procedure has similarities with the fuzzy k-nearest neighbour algorithm described in Section 8.2.5.3.

8.2.2 The single nearest neighbour rule

One of the first descriptions of the nearest neighbour method was provided by Cover and Hart (1967). Under the assumption that points that lie close together in the feature space are likely to belong to the same class, they proposed using a single nearest neighbour rule. This is a special case of the general k-NN rule described in 8.2.1 and involves assigning each point to the class of its single nearest neighbour. A considerable proportion of the literature on nearest neighbour methods has concentrated on this special case. In particular much attention has been focused on the asymptotic convergence of the nearest neighbour classification rule. Cover and Hart (1967) showed that the error rate of the NN rule converges asymptotically to a value between the Bayes misclassification rate r^* and twice this value. This is of interest because it allows one to place bounds on the Bayes error rate. Hand (1981) proves that $r^*(2\text{-}N/(N\text{-}1)r^*)$ is a tighter upper bound on the NN asymptotic misclassification rate r, where N is the sample size. If we did know r^* then we could obtain an indication of the best and worse performance of the NN rule for large samples. In practice this inequality is more useful when inverted to give asymptotic bounds on the unknown Bayes error rate:

$$r^* \geq \frac{N-1}{N} - \sqrt{\frac{N-1}{N}}\sqrt{\frac{N-1}{N} - r}.$$

Because in practice the data analyst may not have access to a suitably large sample for asymptotic results to hold, it is important to understand how the performance of the NN rule will vary with factors such as sample size, dimensionality, metric and distributions. An important question that has been addressed by several researchers concerns the bias of the finite-sample nearest neighbour error from its asymptotic value. Cover (1968) has considered the case of finite sample NN classification rules and found that the bias of the NN error rate from its asymptotic value is bounded by the function $O(N^{-2})$ where N is the sample size. Fukanaga and Hummels (1987) derive an expression for the bias of the NN classifier in terms of the factors mentioned above and show that it can be expressed as the product of two terms, one of which is independent of the distributions and the other of which is independent of the sample size. This allows one to estimate the increase in sample size necessary to reduce the bias by a certain amount, and the relationship between that value and the dimensionality of the data. The results are shown to be a generalization of

Cover's result to more than one dimension. These results are of limited use for our application because we do have access to a large sample, and the size of the sample and the dimensionality of the problem are to a large extent fixed before we begin the analysis.

Another aspect of the NN rule that has not received proportionate attention in the literature is the selection of distance measures. This is a vital factor in the production of successful nearest neighbour and *k*-NN rules and will be considered separately in Section 8.2.3 as it forms a key part of our analysis. For the remainder of this chapter we will concentrate on *k*-NN methods of which the NN method is a special case.

8.2.3 Distance measures for the NN and *k*-NN methods

Fundamental to the identification of the k nearest design points is a distance metric. Much of the work on the NN rule has focused on selecting a distance metric to minimise the difference between the finite sample misclassification rate and the asymptotic misclassification rate. In our case, for commercial reasons, the aim is to select a distance metric to minimise the expected bad rate among the accepts from the classification rule. Ways of doing this are proposed in section 8.3.

We now consider the form of metrics that have been proposed in the literature to measure the distance between a point $\mathbf{x}$ to be classified and a point $\mathbf{y}$ in the training set. Traditionally the most popular measure of distance has been the Euclidean metric given by:

$$d_1(\mathbf{x},\mathbf{y}) = \{(\mathbf{x}-\mathbf{y})^t(\mathbf{x}-\mathbf{y})\}^{1/2}$$

Myles (1991) gives two simple examples to show why the Euclidean metric may not always be the most sensible distance measure to use. The problems arise because the Euclidean metric weights equally the distances parallel to each axis. To try and take account of this the Euclidean metric can be generalised to give:

$$d_2(\mathbf{x},\mathbf{y}) = \{(\mathbf{x}-\mathbf{y})^t\mathbf{A}(\mathbf{x}-\mathbf{y})\}^{1/2}$$

where A can be any p x p symmetric positive definite matrix and p is the dimension of the characteristic space. Note that d_1 occurs when $\mathbf{A} = \mathbf{I}$.

Myles (1991) reviews two approaches to selecting **A**: local metrics are defined to be those for which **A** can vary with **x** and global metrics are defined to be those for which **A** is independent of **x**. In the global case the distance between two points depends only on their relative position in Euclidean space.

8.2.3.1 Local Metrics

Short and Fukunaga (1981) adopt a local metric in order to minimise the distance between the finite sample NN classification error and the asymptotic NN error. If **x** is the point to be classified and $\mathbf{y}_{NN}$ is the nearest neighbour of **x** then the appropriate distance metric is given by:

$$d_3(\mathbf{x}, \mathbf{y}_{NN}) = V^t(\mathbf{x})(\mathbf{x} - \mathbf{y}_{NN})$$

where

$$V(\mathbf{x}) = \frac{d}{dt}|_{t=\mathbf{x}} P(class1|t).$$

This can be derived by considering a linear approximation of $P(class|Y_{NN})$ in a local region of **x**. Because of the linear approximation implicit in the distance metric, we can only apply d_3 to points that are close to **x** in the Euclidean sense. Two methods of using this metric are suggested. The preferred option is to pick a value m and select the m nearest points to **x** using the Euclidean metric. Among these m points the nearest neighbour to **x** is determined by the measure d_3.

Before this metric can be used we need to obtain an estimate of the function V(**x**) described in the metric definition. The authors describe a non-parametric estimate obtained using the mean of the class 1 design points and the mean of all the design points within a local circular region around **x**. A simulation experiment is described in which the above procedure is compared with the Euclidean metric. The authors conclude from this that the new procedure using d_3 is superior to the Euclidean metric d_1, and that the improvement increases as the dimensionality does.

In Fukunaga and Flick (1984) a parametric estimate of V(**x**) is proposed based upon the assumption of a multivariate normal distribution for each class. A simulation experiment is described to compare this metric with the non-parametric version mentioned above and the Euclidean metric. The results

seem to indicate that the parametrically defined metric is the best. However, Myles (1991) raises two important objections. The first was that the criterion for measuring success was closeness to the asymptotic nearest neighbour risk, whereas in practice the metric which gave the smallest error rate would be preferable (this is certainly the case for our particular application). The results showed that the non-parametric version of the method gave a lower error rate when the covariance matrices of the two classes were equal. The second objection was that multivariate normal distributions were used to provide the simulated data making the parametrically defined metric particularly suitable. If the analysis was repeated with non-normal data then we might expect to see a deterioration in performance for this method. As a consequence of these comments it would be dangerous to draw general conclusions about the relative performance of the parametric and non-parametric versions of the metric d_3 from these results.

A local metric has the advantage over global metrics that it preserves the local structure of differences between class membership probabilities ($P(1|\mathbf{x})$ and $P(2|\mathbf{x})$). However, this is achieved at considerable extra computational cost, since the metric itself has to be recalculated for each $\mathbf{x}$. Moreover, a danger of using a local metric approach is that the metric will incorporate local features of the training data set that are not representative of the population from which the sample is drawn (i.e. the estimator will overfit the data). It is also difficult to determine the metric accurately since for each $\mathbf{x}$ the metric has to be determined from a small region around $\mathbf{x}$. For these reasons a global approach to selecting a nearest neighbour metric might be preferred. A global metric is also more likely to be robust to changes in the population. As will be seen in Section 8.3 we choose to take a global approach to metric selection.

8.2.3.2 Global Metrics

Fukanaga and Flick (1984) propose using a global metric that is optimised with respect to the mean-squared error between the asymptotic and finite sample NN risk. They derive an expression for $\mathbf{A}$ that minimises this quantity given by:

$$\mathbf{A}_0 = \int \omega(\mathbf{x})\mathbf{Z}(\mathbf{x})\mathbf{Z}(\mathbf{x})^t P(\mathbf{x})^{1-2/n} d\mathbf{x}$$

where

$$\mathbf{Z}(\mathbf{x}) = P(1|\mathbf{x})P(2|\mathbf{x})\frac{\partial}{\partial \mathbf{x}} P(1|\mathbf{t})$$

The authors present a method of estimating $\mathbf{A}_0$ from a sample of selected points. The metric with this estimated value of $\mathbf{A}_0$ can then be applied to any points in the feature space.

8.2.3.3 A comparison of global and local metrics

Fukanaga and Flick carried out a simulation experiment to compare the global approach with the local metric approach. Four different distance measures are considered: the Euclidean metric (*E*), the non-parametric local distance measure proposed by Short and Fukanaga (*S*), the parametric version (*P*) and the global metric (*A*). Samples of size $N=40n$ (where n is the number of dimensions) were drawn from two multivariate normal distributions where $\|\mu_1 - \mu_2\| = 2.6$ and $\Sigma_1 = \Sigma_2 = \mathbf{I}$. The resulting Bayes error rate was 0.097 and the asymptotic nearest neighbour risk was 0.143. The analysis involved taking 20 test sets from the same distribution and classifying each using the 4 distance measures described. The mean R_X and standard deviation s_X of the misclassification rate was calculated for each measure X. The results are shown in Table 8.2.1.

Asymptotic NN risk		R=0.143							
n	N	R_E	R_A	R_S	R_P	s_E	s_A	s_S	s_P
2	80	0.155	0.162	0.133	0.151	0.0518	0.0642	0.0528	0.0658
5	200	0.161	0.154	0.136	0.143	0.0406	0.0364	0.0287	0.0366
8	320	0.164	0.143	0.135	0.143	0.0197	0.0191	0.0158	0.0246

Table 8.2.1: Sample statistics from Fukanaga and Flick's simulation to compare global and local distance metrics with the Euclidean metric.

The results show that the metric *P* performs the best in terms of closeness to the asymptotic nearest neighbour result. As in the previous study this is unsurprising as the data distributions chosen by the authors give an advantage to this metric. This also applies to the global metric which performs well for $n=8$. We note that, in fact, the lowest misclassification rates are given by the Short and Fukanaga metric *S*.

These results confirm that using a global metric can be a practical way of improving upon the *k*-NN method with the standard Euclidean metric. However, the evidence is limited because the data is simulated and the results

are sensitive to the selected criterion for performance. Fukanaga and Flick (1984) end by acknowledging that the selection of other types of optimal distance measures for NN risk estimation is an area for further study. This provides the motivation for our proposal of a global metric in Section 8.3. Our aim is to extend the problem to the general k-NN case and use the bad rate amongst the accepts as the criterion for optimality. We incorporate into our global measure knowledge of class differences across the feature space.

8.2.3.4 Other variations on the Euclidean metric

The generalisations of the Euclidean metric that we have considered so far have all been of the form

$$d_2(\mathbf{x},\mathbf{y}) = \{(\mathbf{x}-\mathbf{y})^t \mathbf{A}(\mathbf{x}-\mathbf{y})\}^{1/2}$$

A different class of measures, of which the Euclidean metric is one example, arise from the general Minkowski distance:

$$d_4 = [\sum |x_i - y_i|^r]^{1/r}$$

When $r = 2$ we have the Euclidean metric and when $r = 1$ we have another important special case, the city block distance measure:

$$d_5 = \sum |x_i - y_i|$$

One way of looking at distance metric selection is as an attempt to combine different variable types into a single meaningful scale. The city block measure does this in quite an intuitive way according to the sum of differences in units over all the variables. As explained in Chapter 2 we transformed the data in our problem before trying to build a classification rule in order to standardise the scales and provide a suitable ordering of the data. However, the city block measure may still be an effective measure to use and we will include it in our analysis. We will also make a comparison of different values of the parameter r in the general Minkowski distance.

8.2.3.5 Distance measures for categorical data

The distance measures described above are appropriate for continuous data or data in a pseudo-continuous form such as weights of evidence. If we wish to treat the data as discrete dummy variables (taking values of 0 or 1) then it is fairly easy to extend the ideas. Hand (1981) describes one such approach, similar to the city block metric, where the distance between two points is the

number of components in which they differ. A small problem arises because there are $\binom{d}{i}$ cells at a distance i from any given cell. The result is that the *k*th nearest neighbour may be included among several points at the same distance from the point to be classified. This can be solved by taking an appropriate proportion of these points at random.

Extensions of the local and global metric approaches from above to categorical data are considered in Chapter 3 of Myles (1991). Simulation experiments are described which show that the proposed methods can be used to improve on the simple application of the Euclidean metric. We choose not to take this route because the data in weights of evidence form has a natural ordering and has far fewer variables than would be necessary to describe the same data in dummy variable form.

8.2.3.6 Evaluation of the influence of different metrics on *k*-NN performance

Todeschini (1989) describes the results of applying the *k*-NN method to ten data sets using six different metrics and four data transformations. In this discussion we focus on the influence of metrics rather than data transformations, because by using weights of evidence we have already tried to put the characteristic values onto a suitable scale (the selection of optimal data transformations is an area for further research). The following metrics were considered:

(a) Euclidean metric (given by $d_1(\mathbf{x},\mathbf{y})$)

(b) City block metric (given by $d_5(\mathbf{x},\mathbf{y})$)

(c) Correlation (cosine) coefficient: $d(\mathbf{x},\mathbf{y}) = \dfrac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}}$

(d) Camberra metric: $d(\mathbf{x},\mathbf{y}) = \dfrac{1}{n}\sum \dfrac{|x_i - y_i|}{(x_i + y_i)}$

(e) Lance-Williams metric: $d(\mathbf{x},\mathbf{y}) = \dfrac{\sum|x_i - y_i|}{\sum(x_i + y_i)}$

(f) Lagrange metric: $d(\mathbf{x},\mathbf{y}) = \max|x_i - y_i|$

The Camberra (d) and Lance-Williams (e) metrics are not applicable to data with negative values and so are not appropriate for credit data in weights of evidence form. Of the other metrics, we have already discussed the Euclidean (a) and city block (b) metrics. The Lagrange metric (f) is a version of the city block metric where all the emphasis is given to the distance in the direction with the largest component. The correlation coefficient (c) represents a different approach to measuring similarity between points.

The *k*-NN method with the six metrics described above was applied to ten data sets including Fisher's Iris data (see Fisher (1936)). The resulting error rates showed variability between different metrics within data sets. No metric consistently beat the others across data sets and overall metrics (a), (b), (c) and (e) gave similar performance. However, the correlation coefficient (c) and the Lance-Williams metric (e) achieved the lowest average error rates. The Lagrange metric gave the worst average performance.

8.2.4 Selecting a value of *k* in the *k*-NN method

Choice of k is an important issue in the implementation of the *k*-NN method and involves a trade-off between the bias and variance of the probability estimates of $P(g|\mathbf{x})$. As k increases the bias may increase because the points in the training sample being included become further from the point $\mathbf{x}$ to be classified. At the same time the variance of the predicted class membership at $\mathbf{x}$, $Var[\hat{P}(i|\mathbf{x})] = pq/k$ (where p is the probability that a point within the k nearest neighbours belongs to class i), decreases as k rises. Therefore, in order to select a suitable value of k it is sensible to consider a measure that incorporates bias and variance. A common measure that is used is the Mean squared error given by:

$$\begin{aligned} MSE &= E(\hat{P}(i|\mathbf{x}) - P(i|\mathbf{x}))^2 \\ &= E(\hat{P}(i|\mathbf{x}) - E(\hat{P}(i|\mathbf{x})))^2 + (E(\hat{P}(i|\mathbf{x})) - P(i|\mathbf{x}))^2 \end{aligned}$$

$$= \text{var}(\hat{P}(i|\mathbf{x})) + bias(\hat{P}(i|\mathbf{x}))^2$$

This measure will decrease until it reaches an optimal value and then start to increase as the increasing bias outweighs the improvement in the variance (as shown in figure 8.2.1).

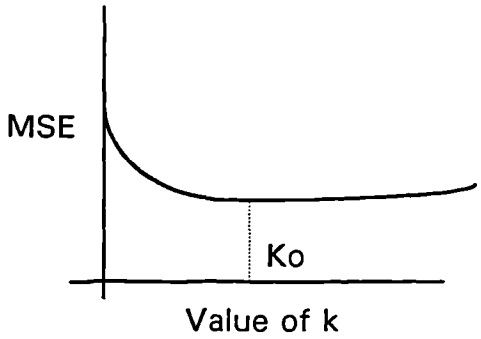


Figure 8.2.1 A plot of Mean Square Error against k.

As has been mentioned before, for our application of the *k*-NN method the objective is to minimise the bad rate amongst the accepted applicants. We are not concerned about the bias and variance of the probability estimates provided that the bad rate is minimised. For this reason the Mean Square Error is not a suitable measurement for us to use to select *k*. Although there is not a direct relationship between the bad rate (or error rate) and the Mean Square Error one would intuitively expect their relationship with *k* to be similar. In fact this has been demonstrated for the error rate for various real data sets as referred to in Myles (1991). Thus we would expect the relationship between bad rate and *k* to be similar to Figure 8.1. We note that the curve of error rate against *k* has often been found to be very shallow near to the optimum value of *k*.

Other work on the selection of an optimal *k* has been carried out by Fukanaga and Hostetler (1973). This includes the development of a functional form for the optimal *k* in terms of the number of points in each class, the sample dimensions and the underlying probability distribution.

Enas and Choi (1986) consider the effects of different covariance matrices and sample sizes on the probabilities of misclassification for different *k* values. They carry out a simulation study using a mixture of continuous and categorical variables. Table 8.2.2 summarises their rough guidelines for a suitable value of *k* to select.

		Difference between sample proportions	
		Small	Large
Difference between covariance matrices	Small	$N^{3/8}$	$N^{2/8}$
	Large	$N^{2/8}$	$N^{3/8}$

Table 8.2.2: Guidelines for selecting optimal k based upon the sample size, N.

The authors also present an adaptive nearest neighbour rule (ANN) which selects k by iteratively maximising the Mahalanobis distance in a local region. They perform simulations to compare this rule with the standard k-NN rule, using their recommended optimal k, and show that the ANN method can be more efficient.

Two points can be made about the results described above with reference to our application. To begin with the data was simulated from particular distributions, including the bivariate normal and Bernoulli distributions. Credit scoring data is categorical and unlikely to fit any standard distribution. A second point to note is that, although the suggested optimal k give quite good performance, they are only crude estimates (based upon the results of one study). If we were to select a value of k for our design set with N = 15054 then the suggested values of k would be 37 and 12. We shall see later in this chapter that these values of k are by no means optimal for our data sets. In order to allow more flexibility in the selection of k, we choose to consider the performance of the k-NN for a broad range of values of k and make selections based upon the corresponding values of the error rate. The issue of selecting a suitable value of k will be returned to in Section 8.6.

8.2.5 Variations of the *k*-NN method

In this section we review other aspects of the k-NN method that have received attention in the literature and may have relevance to our search for the best k-NN classifier for consumer credit data.

8.2.5.1 The reject option

The first variation on the standard k-NN procedure that we will consider is the reject option as introduced by Chow (1957). It uses the basic premise that borderline classifications contribute more significantly to the error rate than cases with more extreme class probabilities. In some situations it may not be necessary to take decisions on the doubtful cases, thus allowing a verdict of no classification to be returned. These cases can then either be ignored altogether or more information can be collected in order to allow a judgement to be made. This choice of whether to classify a point or reject it for classification is called the *reject option.* The result is likely to be a reduction in the error rate. In fact this is very similar to the procedure for vetting credit applicants used by the credit grantor in our problem: a mini scorecard is used to assess all the applicants, leading to a decision to accept or reject 60% of the population (the mini accepts and mini rejects) and defer the decision on the remaining 40%. An application form is then sent to the unclassified applicants leading to more information about the true creditworthiness of the marginal applicants.

For the purpose of evaluating classification procedures we are interested in classifying all applicants and then comparing the resulting error rates. In this study we restricted attention to the construction of a mini-scorecard. As a consequence it is not practical for us to include the reject option. If we were to simulate the two stage classification procedure (with mini and full classifiers) then the reject option could be used to select 40% of applicants at the mini stage for assessment using the full classifier.

8.2.5.2 Distance-weighted nearest neighbour rules

An intuitively appealing adaption of the k-NN rule is described by Dudani (1976) involving giving more weight to observations closer to the point to be classified than to those a larger distance away. The distance-weighted k-NN rule is as follows. Suppose that we wish to classify a point $\mathbf{x}$ and that the k nearest neighbours to $\mathbf{x}$ have distances $d_1 \ldots d_k$. A weight w_i is assigned to each of the neighbours $i = 1 \ldots k$ where

$$w_i = \begin{cases} (d_k - d_i)/(d_k - d_1) & d_k \neq d_1 \\ 1 & d_k = d_1 \end{cases}$$

For each class j we pick out the class j points among the k nearest neighbours to $\mathbf{x}$. The sum of the corresponding weights w_i is then calculated. The classification rule assigns $\mathbf{x}$ to the class for which the weighted sum is greatest.

Dudani performs a simulation experiment to assess the performance of the distance-weighted k-NN rule. The analysis was carried out using three classes and moderate sample sizes. The conclusion drawn is that the distance-weighted rule performs significantly better than the simple majority k-NN rule. He also asserts that the error rate is more likely to remain at a level similar to the minimum as k becomes very large. This would allow one more room for error in choosing a suitable value of k. One of the limitations of the work is that ties were counted as errors and the probability of a tie occurring was much lower for the distance-weighted rule. We note that the distance weighted k-NN rule is similar to the kernel method, where the kernel function weights according to distance from the point to be classified.

8.2.5.3 Fuzzy nearest neighbour rules

Keller et al. (1985) assert that one of the problems with the k-NN decision rule is that each of the points in the training set is given equal importance in determining the class membership of the point to be classified, regardless of how representative of the population that point is. They propose the application of fuzzy set theory to create a fuzzy version of the k-NN algorithm.

Fuzzy sets were introduced by Zadeh (1965) and provide a way of defining categories that are imprecise in nature. This is done through the idea of a membership function which specifies the degree of belonging to a set. We will consider how this is done in the k-NN case. It is assumed that we have a design sample $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n$. A fuzzy c partition of these vectors, which specifies the degree of membership of each class, is given by a $c*n$ matrix $\mathbf{U}$, where $u_{ik} = u_i(\mathbf{x}_k)$ for $i = 1 \ldots c$, and $k = 1 \ldots n$. Here u_{ik} represents the degree of membership of $\mathbf{x}_k$ in class i. The following properties must hold for $\mathbf{U}$ to be a fuzzy c partition

$$\sum_{i=1}^{c} u_{ik} = 1,$$

$$0 < \sum_{k=1}^{m} u_{ik} < n,$$

$$u_{ik} \in [0,1].$$

By selecting suitable values for the class membership of the training set using a fuzzy c partition we are able to tackle the problem that occurs from having "atypical" vectors in the training set. In regions where we are confident that the points in the design set represent the underlying class distributions we can merely use the true class as normal via the crisp partition

$$u_i(\mathbf{x}_j) = \begin{cases} 1 & if \quad \mathbf{x}_j \in class \quad i \\ 0 & otherwise \end{cases}$$

Keller et al. perform an experiment on Fisher's Iris data in which they compare the crisp labelling for the design set defined above with a measure based on the L nearest neighbours of $\mathbf{x}$ given by

$$u_i(\mathbf{x}_j) = \begin{cases} 0.51 + (n_i / L) * 0.49 & if \quad \mathbf{x}_j \in class \quad i \\ (n_i / L) & otherwise \end{cases}$$

where n_i is the number of class i points among the L nearest neighbours to $\mathbf{x}_j$.

The classification rule then involves finding the k nearest neighbours to a point $\mathbf{x}$ to be classified using the Euclidean metric and allocating fuzzy class memberships using the function $u_i(\mathbf{x})$ defined below

$$u_i(\mathbf{x}) = \frac{\sum_{j=1}^{k} u_{ij} \left(1 / \left\| \mathbf{x} - \mathbf{X}_j \right\|^{2/(m-1)}\right)}{\sum_{j=1}^{k} \left(1 / \left\| \mathbf{x} - \mathbf{X}_j \right\|^{2/(m-1)}\right)}$$

where $\{\mathbf{X}_1, \ldots \mathbf{X}_k\}$ represents the k nearest neighbours of $\mathbf{x}$.
The point $\mathbf{x}$ is classified to the class i for which $u_i(\mathbf{x})$ is greatest. Incorporated into the membership function is distance-weighting allowing more emphasis to be given to points near to $\mathbf{x}$ (see above for more discussion of this issue).

The results of the experiment described show that the fuzzy k-nearest neighbour rules can lead to lower misclassification rates than the standard k-NN classifier. The fuzzy k-NN method also allows one to look at the probabilities assigned to incorrectly classifed points in the test set. By doing this it was seen that points $\mathbf{x}$ for which $0.5 < u_i(\mathbf{x}) < 0.7$ were much more likely to be misclassified than points for which $u_i(\mathbf{x}) > 0.7$. Thus the allocated memberships from the fuzzy rule provide a useful measure of confidence in the classifications being made.

In conclusion the fuzzy k-NN rule seems to have potential for application in the credit scoring context. It is able to incorporate the advantages of the distance-weighting method as well as allowing for unrepresentative points in the design set through fuzzy class partitions. Jozwick (1983) has carried out a simulation experiment where a fuzzy k-NN rule performed better then the edited k-NN rule (see below) of Koplowitz and Brown (1981). The ability to measure confidence in the fuzzy rule's predictions is also useful.

8.2.5.4 Methods of reducing the size of the design set

A weakness of the standard k-NN methodology is that the classification of a new point requires the calculation of the distances from that point to all the points in the design sample. This can be computationally slow and means that memory needs to be set aside for storing the design set points. We discuss three modifications of the k-NN method that aim to reduce the size of the design set.

The *condensed-nearest-neighbour* (CNN) rule of Hart (1968) works by splitting the design set into two stores, S_1 and S_2. To begin with all but one of the points in the design set are placed into S_1. The method works by successively taking points from S_1 and classifying them using the NN method with S_2 as the design set. If a point is incorrectly classified it is added to S_2, otherwise it is returned to S_1. The process continues until a complete scan through the remaining points in S_1 finds no new points to transfer. The resulting set S_2 is used as the design set for classification.

The *reduced-nearest-neighbour* rule of Gates (1972) is an extension of the CNN rule. It allows points from S_2 to be returned to S_1 after completion of the CNN procedure. Points are successively removed from S_2 to form a new set $\tilde{S}_2$. Then each point is transferred to S_1 if the particular point and all the points in S_1 are correctly classified using S_2 as a design set. In practice this approach leads to a relatively small reduction in the size of the design set after the CNN rule, given the number of extra iterations needed.

A third approach to reducing the size of the design set is the *edited-nearest-neighbour* rule described by Hand and Batchelor (1978). This involves pre-

processing of the data to remove outliers followed by standard application of the CNN rule. Outliers are removed using the rule:

$$\text{Remove points from class } i \text{ if } \frac{P(i|\mathbf{x})}{P(j|\mathbf{x})} < t, \text{ for some threshold } t.$$

Hand and Batchelor estimate the class membership probabilities using the k-NN method. This innovation is necessary because if outliers are present in the sample to which the CNN method is applied, they are likely to be misclassified by S_2 and, thus, retained in the design set. (Furthermore, neighbouring points from the other class may have to be retained in the design set because otherwise they too will be misclassified by the outlier in S_2.) By using the edited-nearest-neighbour rule the size of the design set can be reduced significantly.

8.3 A proposal for a new approach to metric selection

In Section 8.2.3 we considered distance measures that have been proposed in the literature for the NN method. The major part of the discussion was given to global and local variations of the Euclidean metric. In this section we propose an original k-NN global metric that attempts to incorporate knowledge of class differences contained in the data and show how it can be derived in two ways. The first approach is considered in Section 8.3.1 and centers on selecting a suitable matrix $\mathbf{A}$ for the metric d_2 defined in Section 8.2.3. In Section 8.3.2 we present a more general derivation of the same metric via an appropriate transformation of the data. This can form the framework for a more general approach to metric selection. This framework involves the transformation of the data prior to use of one of the standard metrics available e.g. Euclidean, city block or other form of the general Minkowski distance. In section 8.3.3 we apply this transformation to metrics other than the Euclidean. Other issues relating to the implementation of the methodology are discussed in Sections 8.3.4 and 8.3.5.

8.3.1 The adjusted Euclidean metric

One of the assumptions of the k-NN method is that $P(g|\mathbf{x})$ is constant within the region of the feature space containing the k nearest neighbours. To avoid bias in the probability estimates resulting from the k-NN rule we want to make

this assumption as true as possible. A weakness of the Euclidean metric is that it gives equal weighting to distances parallel to each axis and as a result makes no use of knowledge from the data about how $P(g|\mathbf{x})$ changes with $\mathbf{x}$. To reduce this problem, we define the distance between two points as being the separation between them in the direction orthogonal to equi-probability contours with allowance for random variation. If we did know the equation of the "true" equi-probability contours then the best metric to use would be the distance in the orthogonal direction. In practice the equi-probability contours are estimated using some model form and so we have to make allowance for inaccuracies in this model. The derivation of our metric involves selecting a suitable matrix **A** for the metric

$$d_2(\mathbf{x},\mathbf{y}) = \{(\mathbf{x}-\mathbf{y})^t \mathbf{A}(\mathbf{x}-\mathbf{y})\}^{1/2}$$

Before deriving an expression for **A** we consider a simple example to help illustrate the idea behind our proposal.

Let us assume that we have a sample of data with two characteristics X_1 and X_2 as shown in Figure 8.3.1. Added to the figure are supposed "true" contours of equal $P(g|\mathbf{x})$ and a score line, **w**, is shown in an orthogonal direction.

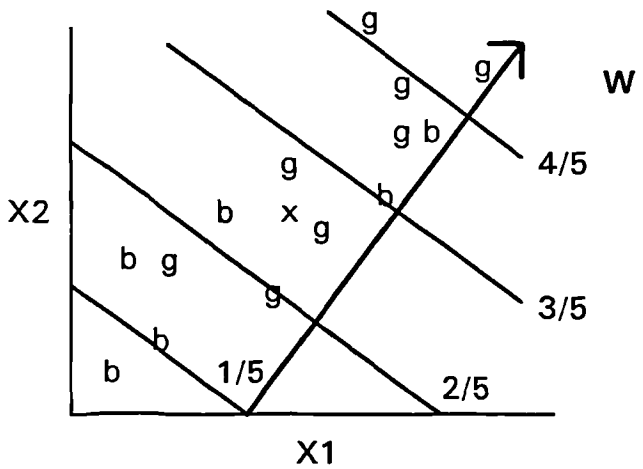


Figure 8.3.1: A good metric to use will give extra emphasis to distance in the direction along the score line, **w**.

In the example shown above, the score direction, **w**, represents the line of maximum class separability and is orthogonal to equi-probability contours. If we know the equation of the score direction then we can use this information to distort the Euclidean distance measure. By giving extra weighting to distance

along the score direction, similarity under our metric corresponds to similarity in terms of $P(class|\mathbf{x})$. This helps to reduce the bias in the probability estimates obtained from the k-NN rule. We would expect this to lead to improved discrimination between classes.

More formally, the squared distance between two points $\mathbf{x}$ and $\mathbf{y}$ in the direction orthogonal to equi-probability contours (denoted by $\mathbf{w}$) is given by:

$$(\mathbf{w}^t(\mathbf{x}-\mathbf{y}))^2 = (\mathbf{x}-\mathbf{y})^t \mathbf{w}\mathbf{w}^t(\mathbf{x}-\mathbf{y}) \qquad (8.1)$$

In practice we estimate the equi-probability contours from the data via a regression model and so our metric needs to take into account random variation in the estimates. Moreover the true contours are unlikely to be exactly linear in practice. (The issue of how to select the equi-probability contours and score direction will be covered in more detail in Section 8.3.4.) For these reasons our proposed metric includes a contribution from the squared Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$ given by:

$$(\mathbf{x}-\mathbf{y})^t(\mathbf{x}-\mathbf{y}) = (\mathbf{x}-\mathbf{y})^t I(\mathbf{x}-\mathbf{y}) \qquad (8.2)$$

Combining terms (8.1) and (8.2) gives the "overall squared distance":

$$(\mathbf{x}-\mathbf{y})^t I(\mathbf{x}-\mathbf{y}) + (\mathbf{x}-\mathbf{y})^t \mathbf{w}\mathbf{w}^t(\mathbf{x}-\mathbf{y})$$
$$(\mathbf{x}-\mathbf{y})^t [I + \mathbf{w}\mathbf{w}^t](\mathbf{x}-\mathbf{y})$$

The measure that we propose comes from the the overall squared distance considered above with a parameter, D, included to allow one to vary the extra weighting given to distance orthogonal to the equi-probability contours. The resulting adjusted Euclidean metric is given by

$$d_6(\mathbf{x},\mathbf{y}) = \{(\mathbf{x}-\mathbf{y})^t [I + D\mathbf{w}\mathbf{w}^t](\mathbf{x}-\mathbf{y})\}^{1/2}$$

Our metric is in the form of the generalisation of the Euclidean metric, d_2, proposed by Fukanaga and Flick (1984) where the matrix $\mathbf{A}$ is given by:

$$A_{D,\mathbf{w}} = (I + D\mathbf{w}\mathbf{w}^t).$$

The metric d_6 includes the distance parameter, D, and selection of a suitable value for this parameter will be an important part of the implementation of the methodology. A simple way to achieve this is to use trial and error: compare the performance of k-NN classifiers using the adjusted Euclidean metric, d_6, for different values of D and select the value which leads to the lowest resulting error rate. An alternative approach is described in Section 8.3.5.

In computational terms the simplest way to view the metric is as

$$d_6(\mathbf{x},\mathbf{y}) = \{(\mathbf{x}-\mathbf{y})^t(\mathbf{x}-\mathbf{y}) + D(\mathbf{w}^t(\mathbf{x}-\mathbf{y}))^2\}^{1/2}$$

We now propose an alternative metric that tries to fulfil the same aim and has a similar form to this version of metric d_6. Our intention in doing this was to explore whether there might be a better way of incorporating knowledge of equi-probability contours from the data. Our proposal for an alternative adjusted Euclidean metric is given by

$$d_7(\mathbf{x},\mathbf{y}) = \{(\mathbf{x}-\mathbf{y})^t(\mathbf{x}-\mathbf{y})\}^{1/2} + D|\mathbf{w}^t(\mathbf{x}-\mathbf{y})|$$

The metric d_7 takes the Euclidean distance and adds to it an extra term that represents a multiple of the distance between two points $\mathbf{x}$ and $\mathbf{y}$ in the direction of $\mathbf{w}$. The difference from d_6 is that the weighted distance in the direction $\mathbf{w}$ is added on as an extra term rather than being incorporated into the Euclidean distance. As a result it is not possible to express d_7 in the general form of metric d_2. Thus there is less theoretical justification for d_7 than d_6, but we include it in our analysis for comparison purposes.

8.3.2 A general transformation of the data

In the last section we derived an adjusted Euclidean metric to take account of information about the distribution of $P(g|\mathbf{x})$ contained in the data. This was done by including in our metric d_6 a term giving extra emphasis to distance orthogonal to equi-probability contours. We now show how this metric can be derived in a more general way through a transformation of the data and the standard application of the Euclidean metric to the transformed data.

The general transformation that we will use is given by:

$$T:\mathbf{x} \mapsto \mathbf{x} + \tilde{D}\mathbf{w}\mathbf{w}^t\mathbf{x}$$

where $\mathbf{w}$ is the direction vector perpendicular to the equi-probability contours and $\tilde{D}$ is a distance parameter (the metric adjustment constant) as before.

Our proposal for a general data dependent k-NN method is to shift all points using the transformation T and then the Euclidean metric (or any other metric) is used to calculate the distance between two points as usual. We can explicitly show the effect of the transformation T on the Euclidean metric and justify that this approach is equivalent to the adjusted Euclidean metric advocated in the previous section.

$$
\begin{aligned}
T{:}d_1(\mathbf{x},\mathbf{y})^2 \equiv T{:}(\mathbf{x}-\mathbf{y})^t(\mathbf{x}-\mathbf{y}) &\mapsto (\mathbf{x}+\tilde{D}\mathbf{w}\mathbf{w}^t\mathbf{x}-\mathbf{y}-\tilde{D}\mathbf{w}\mathbf{w}^t\mathbf{y})^t(\mathbf{x}+\tilde{D}\mathbf{w}\mathbf{w}^t\mathbf{x}-\mathbf{y}-\tilde{D}\mathbf{w}\mathbf{w}^t\mathbf{y}) \\
&= (\mathbf{x}-\mathbf{y}+\tilde{D}\mathbf{w}\mathbf{w}^t(\mathbf{x}-\mathbf{y}))^t(\mathbf{x}-\mathbf{y}+\tilde{D}\mathbf{w}\mathbf{w}^t(\mathbf{x}-\mathbf{y})) \\
&= (\mathbf{x}-\mathbf{y})^t(\mathbf{x}-\mathbf{y})+\tilde{D}^2(\mathbf{x}-\mathbf{y})^t\mathbf{w}\mathbf{w}^t\mathbf{w}\mathbf{w}^t(\mathbf{x}-\mathbf{y}) \\
&\quad +(\mathbf{x}-\mathbf{y})^t\tilde{D}\mathbf{w}\mathbf{w}^t(\mathbf{x}-\mathbf{y})+(\mathbf{x}-\mathbf{y})^t\tilde{D}\mathbf{w}\mathbf{w}^t(\mathbf{x}-\mathbf{y}) \\
&= (\mathbf{x}-\mathbf{y})^t(\mathbf{x}-\mathbf{y})+(\tilde{D}^2\mathbf{w}^t\mathbf{w})(\mathbf{x}-\mathbf{y})^t\mathbf{w}\mathbf{w}^t(\mathbf{x}-\mathbf{y}) \\
&\quad +2\tilde{D}(\mathbf{x}-\mathbf{y})^t\mathbf{w}\mathbf{w}^t(\mathbf{x}-\mathbf{y}) \\
&= (\mathbf{x}-\mathbf{y})^t(\mathbf{x}-\mathbf{y})+D(\mathbf{x}-\mathbf{y})^t\mathbf{w}\mathbf{w}^t(\mathbf{x}-\mathbf{y}) \\
&= d_6(\mathbf{x},\mathbf{y})^2
\end{aligned}
$$

where $D = \tilde{D}^2\mathbf{w}^t\mathbf{w}+2\tilde{D}$.

We have confirmed that applying the transformation T to the data and then using the Euclidean metric is equivalent to using the adjusted Euclidean metric d_6 from above. In fact, as we have mentioned before, this general transformation can be applied to the data before applying any distance metric. In this way all the standard metrics that are used with the k-NN metric can be treated in a data dependent way. In the next section we derive the data dependent versions of other metrics using the transformation T defined above.

8.3.3 Other data dependent metrics

The city block was defined in Section 8.2.3 to be:

$$d_5 = \sum|x_i - y_i|$$

This metric suffers from the same weakness as the Euclidean metric in terms of giving equal weighting to distance parallel to each axis in the feature space regardless of the data. This motivates the suggestion that one could use the general transformation of the data described in the last section to make allowance for class probability differences across the feature space. The effect of the transformation T, from 8.3.2, on the city block metric is shown below.

$$
\begin{aligned}
T{:}d_8(\mathbf{x},\mathbf{y}) \equiv T{:}\sum|x_i - y_i| &\mapsto \sum|(\mathbf{x}+\tilde{D}\mathbf{w}\mathbf{w}^t\mathbf{x})_i - (\mathbf{y}+\tilde{D}\mathbf{w}\mathbf{w}^t\mathbf{y})_i| \\
&= \sum|(x_i - y_i)+\tilde{D}(\mathbf{w}^t\mathbf{x}-\mathbf{w}^t\mathbf{y})w_i|
\end{aligned}
$$

We would expect this data dependent metric to perform better than the standard city block measure.

The data dependent version of the general Minkowski distance can be derived in a similar way and is given by:

$$d_9 = [\sum |(x_i - y_i) + \tilde{D}(\mathbf{w}^t\mathbf{x} - \mathbf{w}^t\mathbf{y})w_i|^r]^{1/r}$$

We will include data dependent versions of the city block measure and the Minkowski distance for different values of r in our analysis described in Section 8.5.

8.3.4 Selection of weights

The general transformation of Section 8.3.2 and the adjusted Euclidean metric of Section 8.3.1 both make use of a vector of weights orthogonal to contours of equal probability. In practice we do not have access to the underlying equi-probability contours for the population of interest. The information we have is in the form of a sample drawn from the underlying population. Before we can make use of the preceding results we need to estimate appropriate contours and the corresponding weights from the data.

As has been mentioned in Section 8.3.1 a simple way of estimating appropriate weights is to fit a linear or logistic regression to the data and use the resulting regression weights. We investigate which of these methods provides the most suitable set of weights. We might expect that the logistic regression weights would best describe the underlying probability distribution and as such would be the most appropriate.

An initial study was carried out using the adjusted Euclidean metrics d_6 and d_7 with linear and logistic weights to determine which set of weights gives the best performance. The criterion for performance of the classification rule was the bad rate at a 70% acceptance rate. We discuss two approaches to comparing the results:

(1) Select optimal values of k from the test set and compare the corresponding bad rates using the two sets of weights for different D. In practice the test set results are unknown until after the accept/reject decision is made and so a value of k would have to be selected from the design set and fixed before classifying the test set. Our simplified approach was taken to allow us to explore the potential of the method, rather than provide a practical classification rule for implementation. By selecting k from the test set predictions, we were able to cut the amount of computation required by a factor of 4.6. This was important

because running the k-NN algorithm for one value of D takes more than two hours. A more realistic method of estimating k and D, which uses the design set, is proposed in Section 8.6. (In fact, we find that the performance of the k-NN method with adjusted Euclidean metrics is fairly insensitive to the choice of k and D.)

(2) Consider graphical plots of bad rate against k for the two sets of weights.

We consider the two approaches in turn:

(1) For each metric and set of weights the procedure was as follows: choose a value of D and find the corresponding value of k which gives the lowest bad rate in the test set (k was allowed to range from 1 to the number of points in the design set); then, repeat the process for a range of values of D. The resulting "lowest" bad rates for each set of weights can then be compared.

The sample used in the comparison is described in Table 8.5.1. Tables 8.3.1 and 8.3.2 show the bad rates for ranges of D values for the two metrics using linear and logistic weights.

Value of D	Linear (bad rate)	Linear (value of k)	Logistic (bad rate)	Logistic (value of k)
1.50	42.73	830	42.79	1230
1.55	42.74	930	42.79	980
1.60	42.73	1000	42.76	1630
1.70	42.74	990	42.74	1170
1.75	42.74	920	42.77	1160
2.00	42.76	1560	42.80	1040

Table 8.3.1: Bad rates for the k-NN classifier with the adjusted Euclidean metric d_6 using linear and logistic weights.

Value of D	Linear (bad rate)	Linear (value of k)	Logistic (bad rate)	Logistic (value of k)
0.25	42.81	860	42.79	1000
0.30	42.75	870	42.78	960
0.35	42.67	930	42.75	1060
0.40	42.72	970	42.78	1110
0.50	42.75	820	42.86	760
0.60	42.85	750	42.86	140

Table 8.3.2: Bad rates for the k-NN classifier with the adjusted Euclidean metric d_7 using linear and logistic weights.

The results are only presented for values of D around the minima. They show that the linear and logistic weights lead to similar optimum performance of the k-NN method with adjusted Euclidean metrics. We anticipated that this would be the case because the results of Chapter 7 showed that linear and logistic regression give similar ratios between model coefficients for our sample. However, a few minor points can be made about the results:

- The minimum bad rate for the linear weights is flatter than the minimum for the logistic weights when metric d_6 is used.

- In general the linear weights have a lower "optimal" k for both metrics.

- The metric d_6 is less sensitive to changes in D than metric d_7.

- Although the results are similar for the two sets of weights, the linear weights give bad rates that are consistently just below the corresponding logistic weights. The average bad rates (for the ranges of D values considered) are shown in Table 8.3.3:

Weights	Metric d_6	d_7
Linear	42.74	42.76
Logistic	42.78	42.81

Table 8.3.3: Averaged bad rates for linear and logistic weights using metrics d_6 and d_7.

(2) Graphical comparisons of performance using linear and logistic weights:

To obtain a more global picture of the relative performance of the k-NN classifiers with linear and logistic weights, we considered plots of bad rate against k for the two sets of weights. Figures 8.3.2 and 8.3.3 show examples of these curves for the test set with the optimal D.

The curves illustrate the similarity of the two classifiers. In order to make any differences more apparent, Figures 8.3.4 and 8.3.5 show the corresponding difference in bad rates between the two classifiers for different values of k (bad rate for logistic weighted classifier - bad rate for linear weighted classifier). These show the marginally superior performance of the k-NN classifier with linear weights for most k. It is interesting to note how the classifier with logistic weights suddenly starts to perform better than the classifier with linear weights for k near to 3000. This is consistent with point (b) from above, where it was mentioned that the optimal k is higher when logistic weights are used. When considering the curves in the above figures, it should be remembered that the differences are small (all with a magnitude of less than 0.001, which corresponds to approximately 4 applicants).

To end this graphical comparison, Figures 8.3.6 and 8.3.7 show the difference in bad rates for the linear and logistic weighted k-NN classifiers for the range of k and D values together using metrics d_6 and d_7. These plots display an undulating bad rate surface. The curves for particular D, such as Figure 8.3.5, seem to display a periodic relationship similar to a sine wave. This indicates that the relative performance of the linear and logistic weighted classifiers is sensitive to changes in k. Figures 8.3.6 and 8.3.7 also indicate some sensitivity

Fig 8.3.2: Comparison of linear and logistic weights (metric d6, D = 1.55)

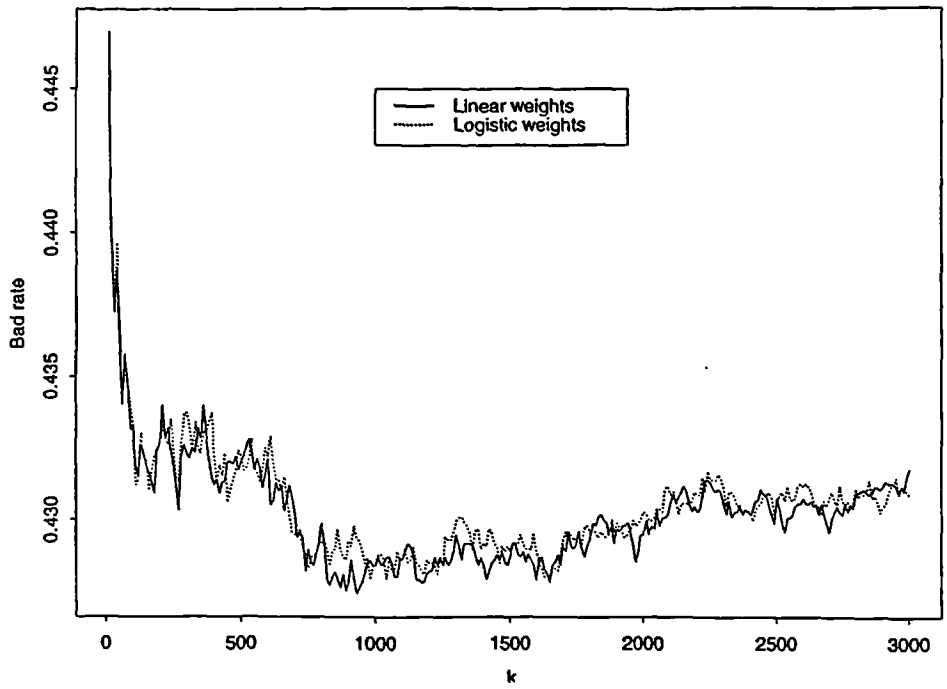


Fig 8.3.3: Comparison of linear and logistic weights (metric d7, D = 0.35)

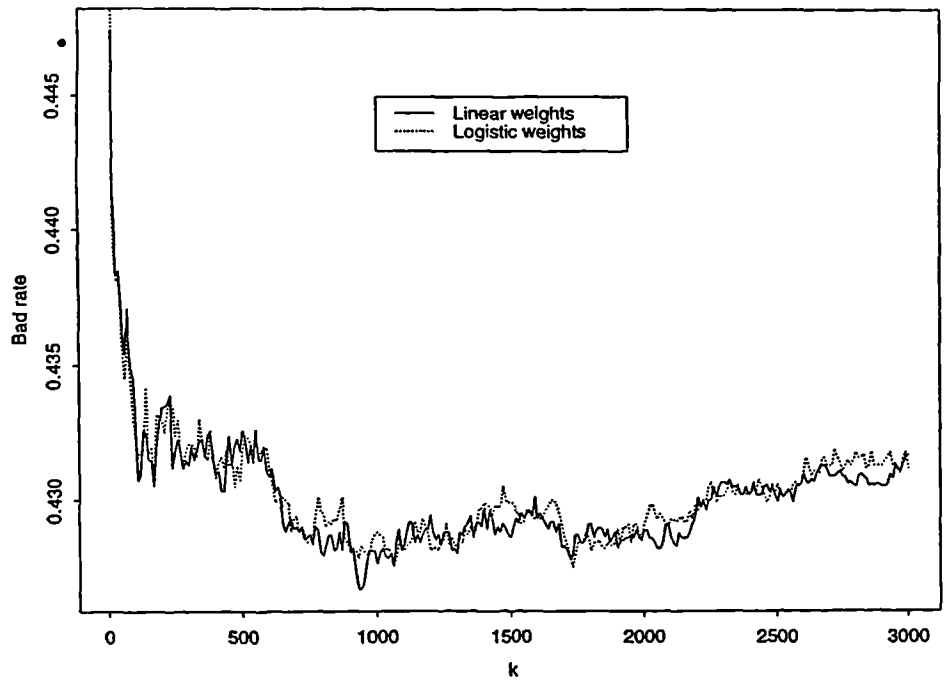


Fig 8.3.4: Logistic - Linear bad rate for metric d6, D = 1.55

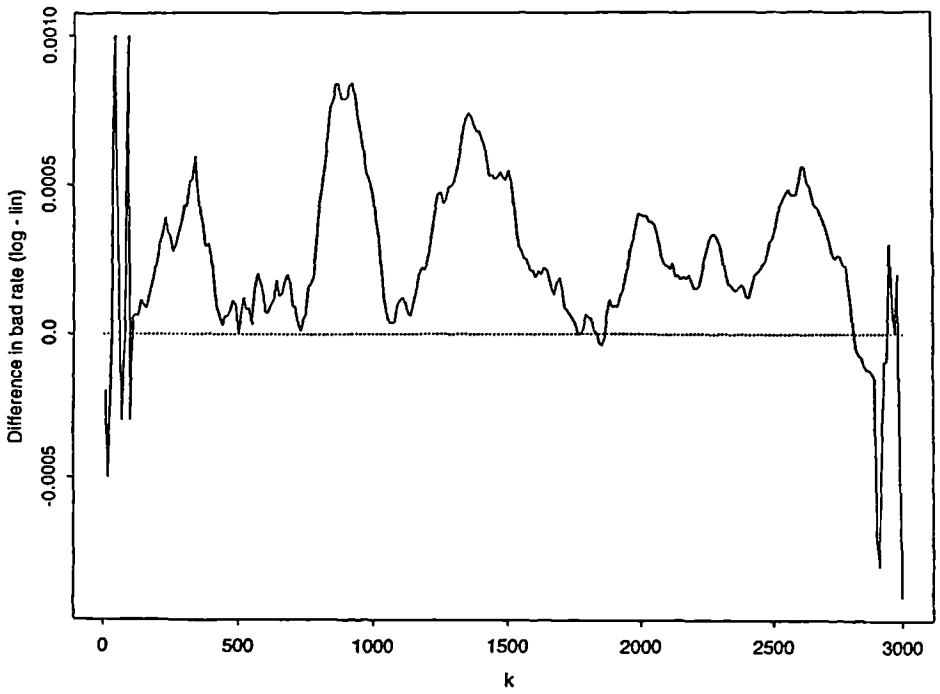


Fig 8.3.5: Logistic - Linear bad rate for metric d7, D = 0.35

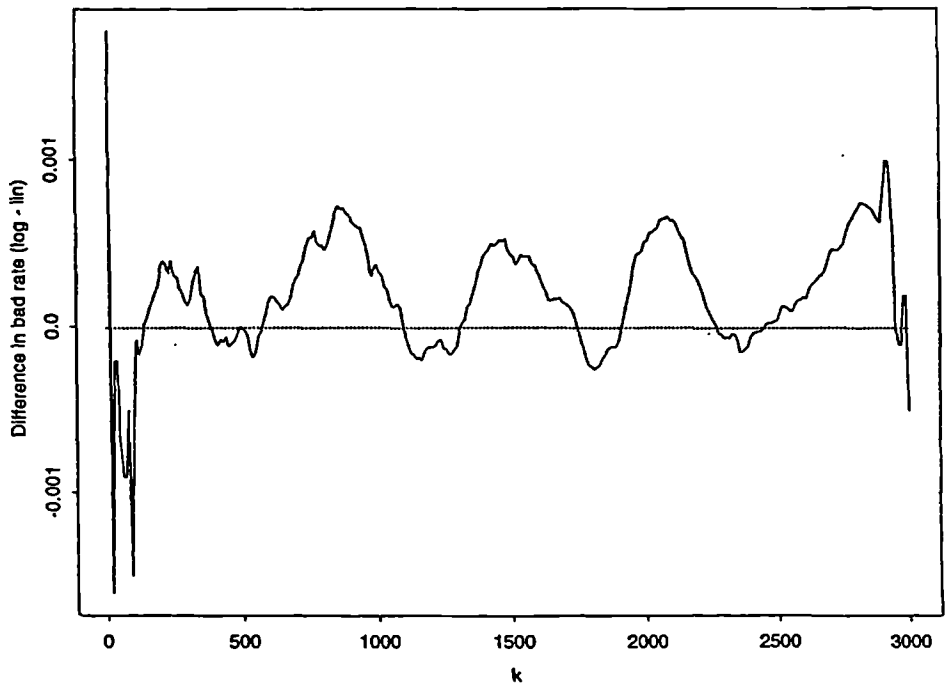


Fig 8.3.6: Difference in bad rate between k-NN with logistic and linear weights (metric d6)

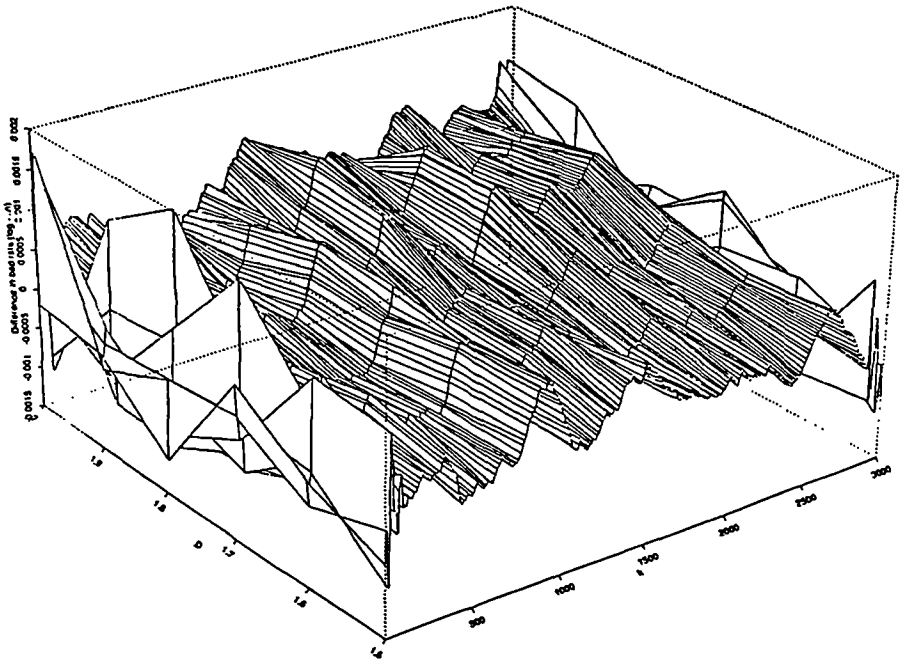
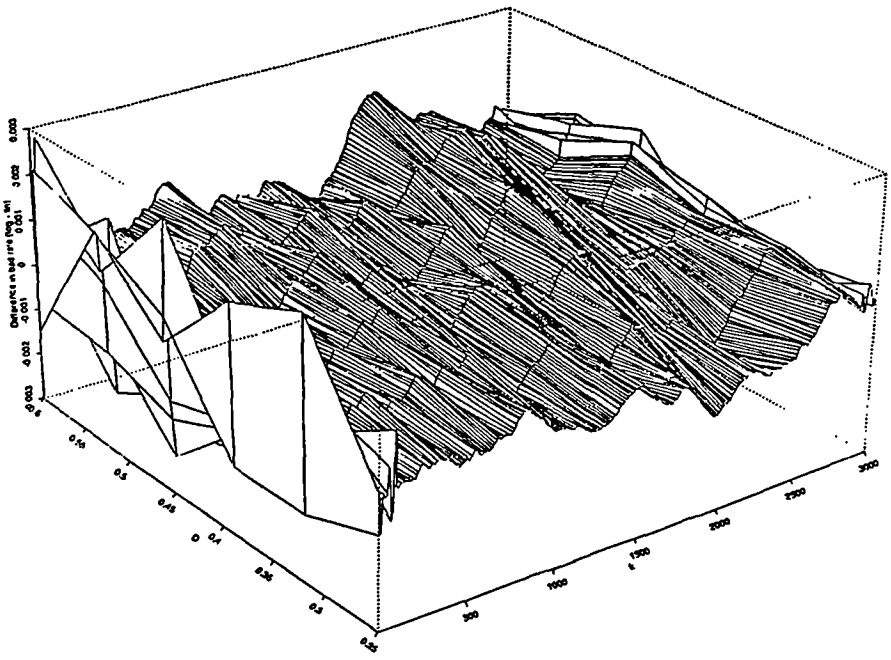


Fig 8.3.7: Difference in bad rate between k-NN with logistic and linear weights (metric d7)



of the bad rate differences to changes in D: the logistic weights give relatively better performance for low values of D (more noticeable for metric d_7).

The implication of the above plots is that the choice of weights should depend upon the estimated parameters k and D. Because the k-NN classifier with linear weights performs at least as well as the k-NN classifier with logistic weights over the range of optimal k and D, we choose to focus on linear weights in the rest of this chapter. However, further work is needed on the identification of the most appropriate weights to use. Until now the models that we have used to describe the class probability structures in the data have involved linear probability contours. We could consider more complex model forms and in particular quadratic probability contours might lead to a better description of the data structure.

8.3.5 A variable distance parameter for the adjusted Euclidean metric

The approach to metric selection that we have proposed in this chapter has stressed the need to take into account knowledge from the data about how $P(g|\mathbf{x})$ changes with $\mathbf{x}$. This involved a transformation of the data which gave extra weighting to the distance along a specified direction vector. The amount of this weighting can be controlled through the distance parameter, D. One way to select an optimum value for D is to perform an exhaustive search through all possible values.

An alternative method of selecting D is to let the emphasis given to distance in the direction orthogonal to equi-probability contours depend on the rate of change of $P(g|\mathbf{x})$. One way of doing this is to make the distance parameter, D, equal to a multiple of the slope of the score direction, $\mathbf{w}$.

If linear regression weights are used then the slope of the score direction is constant and so this has no effect on the value of D selected. In the case of logistic regression the score direction does have a variable slope. If the predicted probability of belonging to class 1 from the logistic regression is

$$P = e^{\beta'\mathbf{x}} / (1+e^{\beta'\mathbf{x}})$$

where β is a vector of regression parameters, then we set

$$D(\beta^t \mathbf{x}) = c\frac{\partial P}{\partial(\beta^t \mathbf{x})} = ce^{\beta^t \mathbf{x}} / (1+e^{\beta^t \mathbf{x}})^2$$

where c is a scaling constant. We note that it is still necessary to select a suitable value for the constant c defined above.

We did some initial investigations to test whether this approach could improve upon the adjusted Euclidean metric with constant D. Initial results are presented below for different values of the parameter c.

Value of c	Lowest bad rate	Value of k
1	42.92	1000
2	42.79	1140

Table 8.3.4: Results of using a variable distance parameter with the adjusted Euclidean metric d_6 and logistic weights.

Table 8.3.4 shows the lowest raw bad rates that were obtained, and the corresponding values of k, for each value of the parameter c. These results are similar to those obtained using a fixed distance parameter and suggest that there is no advantage to be gained from adopting this approach. We hypothesise that this is because there is an optimal value for the parameter D across the entire characteristic space (see Section 8.5.1) and by varying D we end up using sub-optimal values in some regions. As a result the variable distance parameter will give bad rates close to or worse than the value obtained from the optimal fixed D. We focus on using the adjusted Euclidean metric with linear weights and fixed D in the rest of this chapter.

8.4. The implementation of the *k*-NN method with adjusted metrics

Having described the theoretical basis for our approach to metric selection for the k-NN classifier, we implemented the proposed method and carried out experiments to assess its performance. The experiments involved comparisons of our k-NN classifier with other standard classification techniques applied to two credit scoring data sets. These comparisons are covered in detail in

Sections 8.5, 8.6 and 8.7. In this section we restrict attention to practical considerations relating to the implementation of the *k*-NN method.

We begin by highlighting a fundamental practical difference between the *k*-NN method and the traditional method of scorecard building using a technique such as linear or logistic regression. As described in Chapter 2, the traditional method involves assigning numerical values to each of the attributes of the relevant characteristics. A future applicant receives an overall score by summing his/her particular attribute scores, which is then compared with an overall threshold score. The scorecard consists of a list of characteristics with the appropriate attribute scores created by the statistical model.

In the case of the *k*-NN method we cannot build a scorecard with weights for each variable. To classify a future applicant we have to calculate the distance from the new point to all the points in the existing design set and then select the *k* nearest neighbours and calculate a good/bad probability estimate based upon the proportions of goods amongst them. There are two disadvantages of this approach: first, it is computationally expensive to calculate these distances for each new applicant that we wish to classify; secondly, we are not able to specify the contribution that each characteristic makes to the accept/reject decision in a meaningful way (it is arguable whether we are able to do this when building a traditional scorecard).

Despite the drawbacks of the *k*-NN method highlighted above, it was our hope that this approach would lead to improved discrimination between good and bad applicants. We implemented the method with data dependent metrics on a Sun workstation using a "C" program. There were two main problems encountered during the implementation of the method and we will mention each in turn.

Problem 1

First we found that our method of storing and updating the *k* nearest neighbours was very inefficient. This was based upon maintaining an array of the *k* nearest neighbours to date, then calculating the distance to the next point and shuffling it up the table one-by-one until it reaches a point that has a smaller distance. We were able to decrease the running time of the program by about 40 times using a tree-sorting algorithm.

Problem 2

A second major problem that was encountered while implementing the k-NN method concerned the interpolation function that was used to compare the results. As has been described in Section 8.2.1 and above, the classification rule involves estimating a probability of belonging to each class for the applicant under consideration and comparing this with a threshold. This threshold is chosen to allow a fixed proportion of the test set to be accepted and is calculated by rank ordering the estimated good/bad probabilities for the test set (this is unusual for a classification problem- see Chapter 5 for a detailed discussion of the assessment of scorecards). To compare different classifiers we fix an acceptance threshold that we want to compare at (e.g. 70% of applicants) and interpolate the bad rates from the nearest possible acceptance rates to this threshold for each classifier.

A problem that arose was that, for very low and high values of k, there was only a small range of possible estimated $P(g|\mathbf{x})$. The result of this was that the nearest possible acceptance rates to 70% were some distance away. This led to highly inaccurate predictions of the bad rate at the fixed threshold using our initial interpolation function. This was important because we might expect the k-NN classifier to perform well for low values of k, thus necessitating accurate estimates of the bad rate.

The initial interpolation function was given by:

$$I_1 = \{(b_1 * d_2) + (b_2 * d_1)\} / (d_1 + d_2)$$

where b_1 and b_2 are the bad rates corresponding to the acceptance rates above and below the threshold and d_1 and d_2 are the differences between these acceptance rates and the threshold.

We devised a new interpolation function that took account of the number of applicants above and below the threshold. It is given by:

$$I_2 = \{(n_1 * b_1 * d_2) + (n_2 * b_2 * d_1)\} / (n_2 * d_1 + n_1 * d_2)$$

where n_1 and n_2 are the numbers of applicants above the corresponding acceptance rate.

This problem is neatly illustrated by considering a different data set introduced by Buntine and Niblett (1992). The data comes from a study of pole balancing among human subjects and a summary of its properties is given in Table 8.4.1.

	Continuous Attributes	Classes	Number of subjects	% in class 1
Pole data	4	2	1847	51

Table 8.4.1: The pole balancing data set supplied by Buntine and Niblett (1992).

The data set described above was randomly split into a design set of 1000 cases and a test set of 847 cases. The test set was then classified using a k-NN rule based on the design set and the resulting error rate among the cases from class 1 was calculated (this is analogous to our usual criterion of looking at the bad rate among the accepted applicants). The curve of error rate against k for the old interpolation function is shown in Figure 8.4.1.

The lowest value of the error rate at an acceptance rate of 50% is 0.2654 when $k = 997$. If we were to trust this value then we would always choose to look at 997 out of the 1000 cases in the design set in order to make a classification on a new case. This is ridiculous and is a result of the original interpolation function. When the new interpolation function is added in Figure 8.4.2 we see that for k = 997 the error rate is 0.4770. This value is consistent with the error rates for similar values of k and reflects the expected poor performance as k becomes close to the number of cases in the design set.

The example we have just presented shows how the original interpolation function used can lead to highly erratic and inaccurate error rate predictions for high and low k. We have solved this problem by introducing an interpolation function that takes into account the number of cases contributing to a particular error estimate. This new interpolation function has been incorporated into the working version of the program.

Fig 8.4.1: Bad rate curve for pole data with old interpolation function

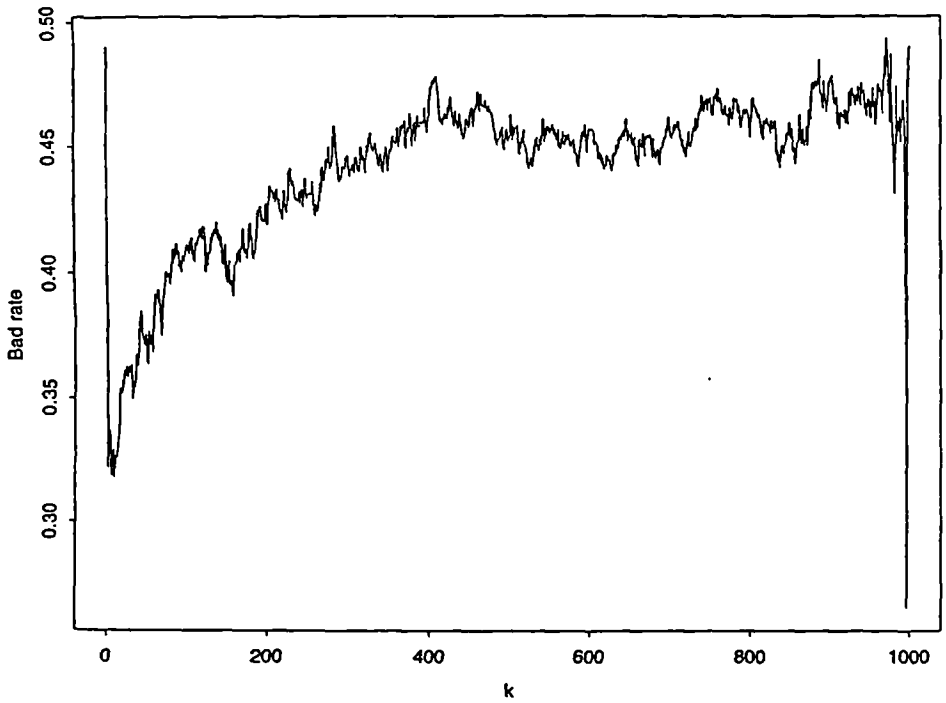
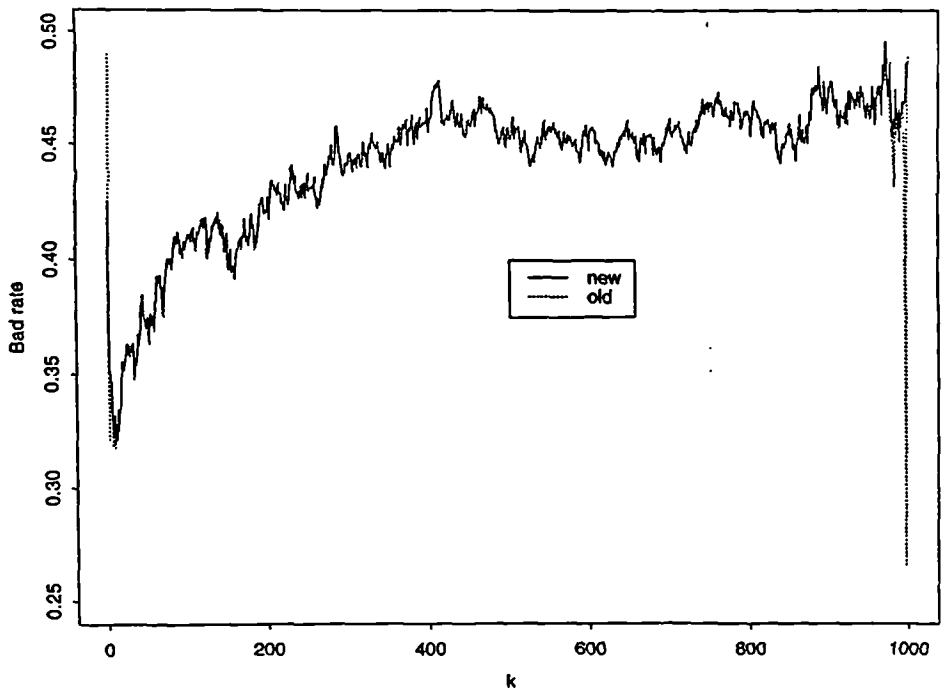


Fig 8.4.2: Bad rate curve for pole data with new interpolation function



8.5 An investigation into the properties of the *k*-NN classifier

This section describes the results of a first investigation into the properties of the *k*-NN classifier applied to consumer credit data. The emphasis is on the potential of the methodology, rather than the practical implementation. For this first study we choose to select k and D using the oversimplified approach described in Section 8.3.4: namely, by taking the values of the parameters which give the lowest bad rates for the test set. This gives us an upper bound on the performance of our method, allowing us to evaluate whether it has the potential to improve upon other techniques. The results for linear and logistic regression and decision trees are presented in order to provide a baseline for the *k*-NN method to beat. We also explore the relationship between bad rate and k. [A more extensive analysis is carried out in Section 8.6, including the proposal of a practical method for selecting the *k*-NN model parameters (k and D) using the design set (this version of the *k*-NN rule can be implemented in practice). A more robust comparison of the different classifiers is also presented.]

The data set used in this study is an updated version of the sample described in Table 2.1. Summarising statistics are given in Table 8.5.1.

	Number of variables	Number of classes	Number of cases	% of bads in full sample
Design set	16	2	15054	54.49
Test set	16	2	4132	54.70

Table 8.5.1: A description of the data set used for assessing the performance of the *k*-NN classifier.

One interesting feature to note is that there are only two classes (good and bad). We choose to exclude the *others* from the analysis in order to concentrate on discrimination between goods and bads. The data is used in weights of evidence form (see Chapter 2). Characteristics were selected on the basis of availability at the mini vetting stage and an examination of information values. The same set of characteristics is used for each of the techniques considered in this chapter.

The set of characteristics included two decision trees that were used to take account of interactions between characteristics. This was done to give assistance to the techniques using assumed linear relationships (linear and logistic regression). This approach is similar to the hybrid classifier of discriminant analysis and decision trees proposed by Boyle et al. (1992). We would expect that by adopting this approach we would remove one of the advantages of the k-NN method and thus set a harder baseline for it to beat.

Performance is assessed using the bad rate amongst the accepts (see Chapter 5 for a discussion of this criterion). The bounds on classifier performance come from putting $p = 45.3\%$ and $a = 0.70$ into the appropriate equations in Section 5.2.2 and are shown in Table 8.5.2.

Description	Bad rate
Best	35.3%
Random	54.7%
Worst	78.1%

Table 8.5.2: Bounds on the bad rate.

8.5.1 Properties of the bad rate curves for the adjusted Euclidean metrics

Figures 8.5.1 to 8.5.4 show plots of bad rate amongst the test set accepts against k for a combination of different metrics and values of D as described below. They illustrate the different types of bad rate curves encountered in practice.

(a) The standard Euclidean metric ($D = 0$):

Fig. 8.5.1: $0 < k < 15050$

Fig. 8.5.2: $0 < k < 3000$

(b) The adjusted Euclidean metric d_6:

Fig. 8.5.3: $0 < k < 3000$ and a range of D values from the interval $0.5 < D < 5$

Fig 8.5.1: k-NN with D = 0 (and 0 < k < 15050)

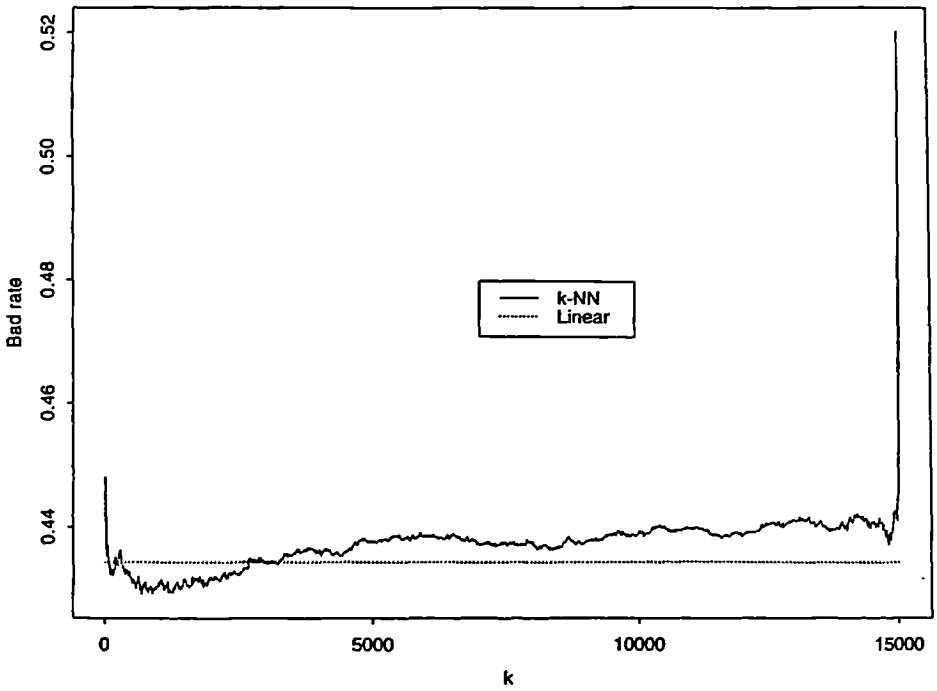


Fig 8.5.2: k-NN with D = 0 (and 0 < k < 3000)

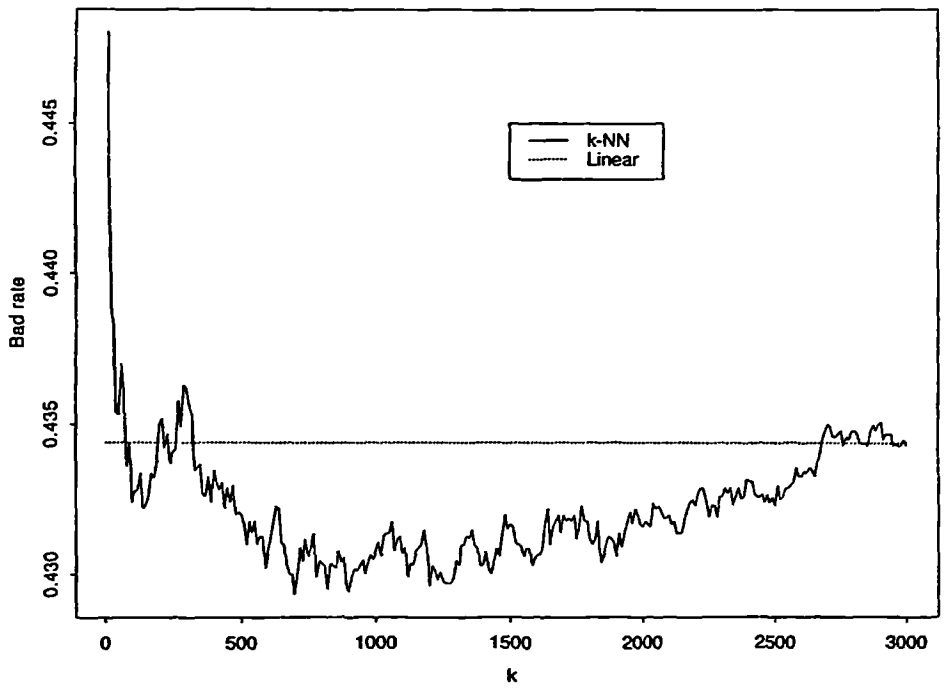


Fig 8.5.3: Bad rate curves for metric d_6

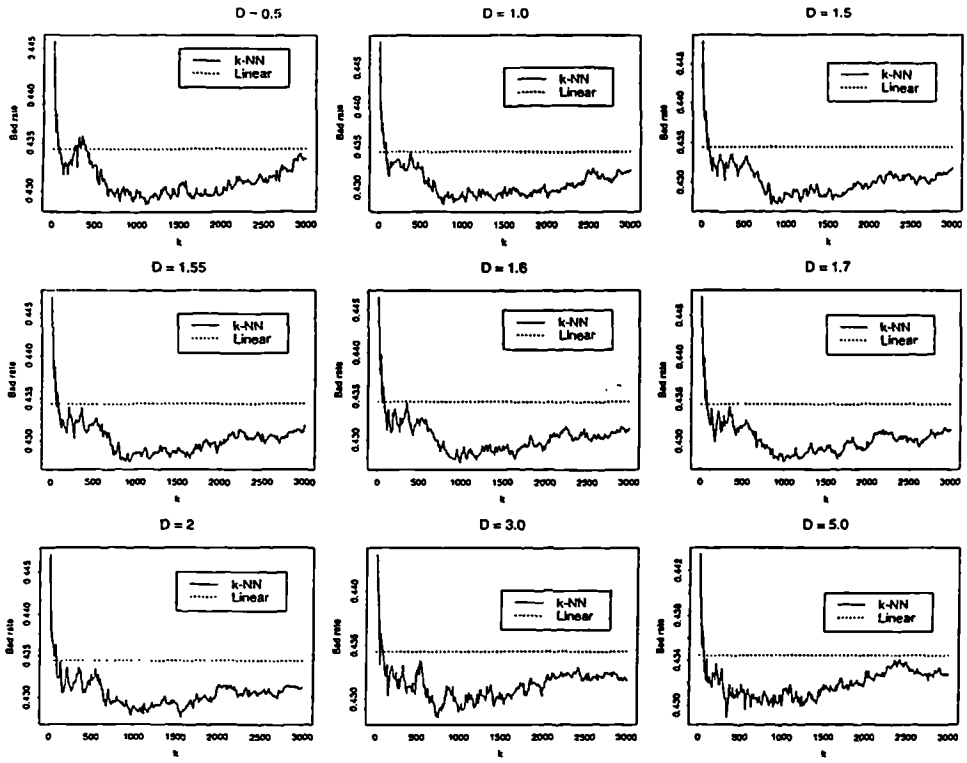
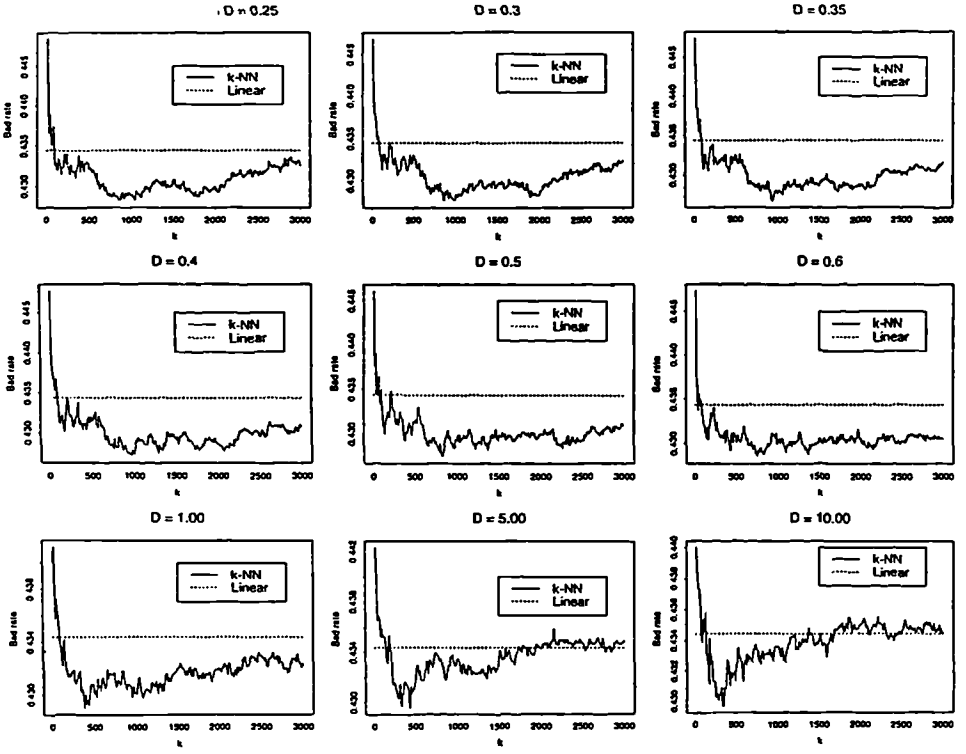


Fig 8.5.4: Bad rate curves for metric d_7.



(c) The adjusted Euclidean metric d_7:

Fig 8.5.4: $0 < k < 3000$ and a range of D values from the interval $0.25 < D < 10$

The horizontal broken lines represent the performance of a classifier using linear regression. This provides a baseline against which to compare the performance of our k-NN classifier. A number of observations may be made about the above figures:

(i) The curves appear very jagged. In fact, the bad rate criterion was evaluated for k=10, 20, 30, ... up to the indicated maximum values. If it had been evaluated for every value of k the curves would have been even more irregular. Marked jaggedness clearly has implications for choosing a value of k, since it implies that slight differences will produce large consequences. However, examination of the vertical axis shows that even the largest jumps in the body of the curve are produced by no more than about ten data points. That is, although the curve appears irregular, the consequences are fairly small. Despite this, we propose a smoothed version of the k-NN estimator in Section 8.6.1.2.

(ii) The early parts of the curves are more jagged - even less regular - than the later parts. These curves are the end product of two averaging processes. First, the averaging inherent in using the k nearest neighbours to estimate probabilities. And second, the averaging over test set points to produce an overall estimate of performance at each particular value of k. For the test set this second part is always an average of 4132 estimated probabilities but the first part involves averaging over samples of size k. Since early parts of the curve are based on smaller k values the associated probability estimates will be expected to have larger variances (the variance of the estimated probabilities from the k-NN rule is given by $Var[\hat{P}(i|\mathbf{x})] = pq / k$ where i is the class, p is the true probability of belonging to class i and $q = 1 - p$).

There may also be a dependence effect since, the larger the value of k, the more design set points each classification is likely to have in common. (With k equal to the entire design set all classifications would be based on the same data

set and so the curve would rise up to the full sample bad rate. This is illustrated by Figure 8.5.1)

(iii) The k-NN curves remain below the linear regression line for a considerable range of values of k and, in general, the range of "best" values of k is large - the curves have broad flat valleys. The breadth of these valleys came as something of a surprise to us. We might reasonably expect that as k is increased the rising bias of the probability estimates will overcome any improvement in the variance of the estimates, resulting in a consequent loss of classification accuracy.

In order to explore this phenomenon we considered plots of $\hat{P}(g|\mathbf{x},k_1)$ against $\hat{P}(g|\mathbf{x},k_2)$ for the points in the original test set, using different values of k_1 and k_2. Figure 8.5.5 shows such a plot when $k_1 = 100$ and $k_2 = 1000$. The increased variation in the ordering of points in the middle of the range is due to the higher variance of the probability estimates in this region. The solid lines represent the 70% thresholds for each classifier.

Figure 8.5.5 shows that, although there is variation in the ordering of the test set points using the k-NN method with the two k values, most points are placed into the same class by the two classifiers. This means that the points accepted using both k values swamp the points only accepted using one.

Table 8.5.4 shows the numbers of good and bad applicants accepted by the k-NN classifier using either one or both of $k_1 = 100$ and $k_2 = 1000$.

	Goods	Bads
Accepts (k = 100), Rejects (k = 1000)	49	123
Accepts (k = 1000), Rejects (k = 100)	53	119
Accepts (k =100 and 1000)	1594	1127

Table 8.5.4: Numbers of goods and bads accepted using the k-NN method with k = 100 and/or 1000.

The proportions of bads amongst the accepts in rows 1 and 2 are very similar, indicating the similar creditworthiness of the swapsets. This, combined with

Figure 8.5.5: P(goodlx) plotted for k = 100 against k = 1000

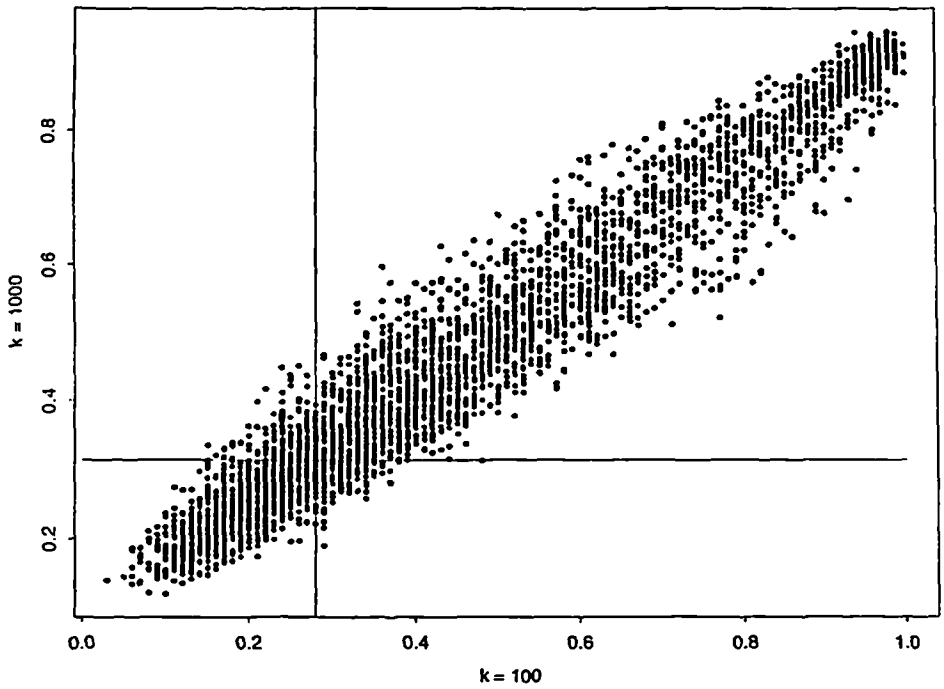
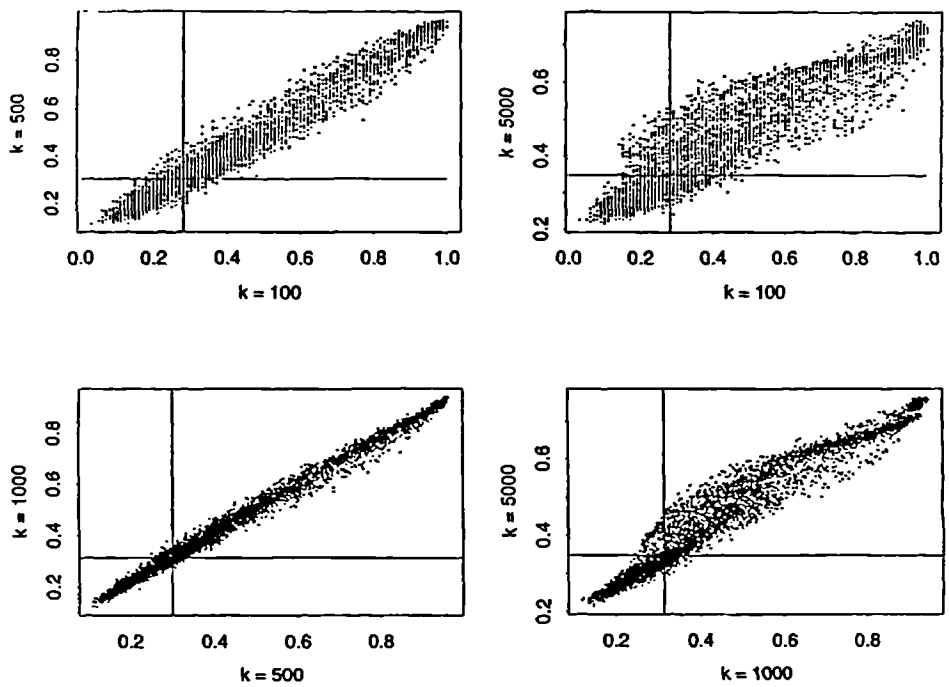


Fig 8.5.6: $P(g|\mathbf{x})$ plotted for different values of k.



the "swamping" effect of the points accepted under both classifiers, helps to explain the "flat valley" phenomenon.

Figure 8.5.6 shows further plots of $\hat{P}(g|\mathbf{x},k_1)$ against $\hat{P}(g|\mathbf{x},k_2)$ using different values of k_1 and k_2. The plots show how the variation in the ordering of points increases as the difference in k values increases. This is not surprising because a larger change in k corresponds to an equal or greater number of "uncommon" points used to make each classification. In each plot the "swamping effect" described above can be observed.

(iv) The consequence of the flat valley phenomenon described above is that the bad rate stays near to its minimum until the value of k represents well over 90% of the design sample (for example see Figure 8.5.1). This indicates that if we merely exclude the points that are most different from the point to be classified then we will get a reasonable prediction of the true class. This is a sort of "most distant neighbours" approach. This feature of the results is unexpected. We might reasonably expect that as k is increased a rising bias of the probability estimates will overcome any improvement in the variance of the estimates resulting in a consequent loss of classification accuracy (see Section 8.2.4).

(v) The implication of the flat minimum described above for choosing an optimal value of k is that there is a wide choice of suitable values of k. In this section the optimal k is chosen such that it gives the lowest bad rate in the test set (it represents an upper bound on the performance that can be achieved in practice for the test set). The optimal k for the adjusted Euclidean metric, d_6, are shown for different D under this criterion in Table 8.5.5.

It can be seen from Table 8.5.5 that there is considerable fluctuation in the selected value of k for different values of D (the selected k range from 340 to 1560). This is not a cause for concern because the flat minimum of the bad rate curve means that we expect good performance over a wide range of k.

Value of D	Value of k	Bad rate
0.00	700	42.93
0.50	1110	42.86
0.75	770	42.85
1.00	780	42.79
1.50	830	42.73
1.55	930	42.73
1.60	1000	42.73
1.70	990	42.74
1.75	920	42.74
2.00	1560	42.76
2.50	740	42.82
3.00	750	42.83
5.00	340	42.89
100.00	Any	43.40

Table 8.5.5: Bad rates at a 70% acceptance rate for metric d_6 with different values of D.

The values for the optimal k (all over 340) shown in Table 8.5.5 are a lot higher than are normally used in applications of the k-NN method (of course, we have a relatively large design sample so we would expect a relatively higher k). In particular our optimal k are higher than the values proposed by Enas and Choi (1986) and discussed in Section 8.2.4. For $N = 15054$, the proposed values of k were 37 and 12 depending upon the covariance structure and the proportions of each class.

In Section 8.5.4 we take a graphical approach to estimating the bias for different values of k and propose an explanation for the high optimal k: we believe that the structure of the population is such that $P(g|\mathbf{x})$ lies between 0.2 and 0.8 for a higher than expected region of the characteristic space. This implies that bias might be expected to increase relatively more slowly than variance decreases, resulting in a high optimal k.

(vi) Throughout this initial study, the linear regression bad rate of 43.44 is used as the baseline against which to compare the k-NN results. It has been added to each of Figures 8.5.1 to 8.5.4. It can be seen that the k-NN method

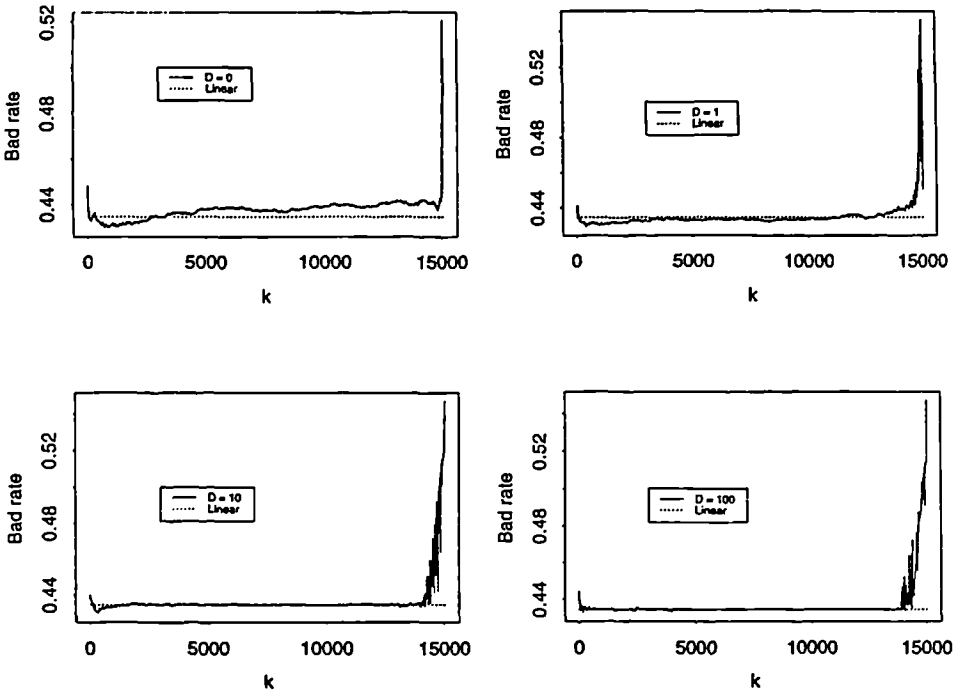
with adjusted Euclidean metrics outperforms linear regression for k between about 300 and 2500 as D varies.

(vii) Figure 8.5.7 shows what happen to the bad rate curves as D increases to a large value (in this case 100) using metric d_6. Such a large D means that distance is effectively being measured orthogonally to the equiprobability contours of $P(g|\mathbf{x})$. As a result the characteristic space is reduced to a one dimensional score line and the k-NN method involves looking at differences in score. Because the method used to provide the weights (linear regression in this case) gives a direction of increasing $P(g)$, the k-NN method will tend to keep the same ordering of the test set as comes from the regression. This means that the method of selecting weights and the k-NN method with an adjusted Euclidean metric will give very similar performance (although the actual estimated probabilities may differ). The figures illustrate how the curves "converge" to the linear regression line as D increases. It can be seen that the curve has converged to the linear regression line for most k by $D = 10$. The only appreciable difference between the curves for $D = 10$ and $D = 100$ is the region of low k (i.e. the optimal region where $k < 1000$).

Another feature of Figure 8.5.7 is the large fluctuations in bad rate for high k. From (i) we know that the bad rate curve must reach the sample bad rate as k tends to the number of points in the design set. From above we know that for high D the k-NN method gives similar performance to linear regression. As k becomes very large the influence of the first of these factors starts to overcome the second. The large fluctuations are due to the instability of the probability estimates for very high k.

To complement the above plots of k, Figures 8.5.8 and 8.5.9 show curves of bad rate against D for the test sample using metrics d_6 and d_7. The curves show global minima, as we hoped, although the differences in bad rate are quite small. We conclude that the bad rate is fairly insensitive to the choice of D. The only difference between the curves is that minimum valley is slightly broader for metric d_6. The irregular shape of the curves is due to the limited number of D values considered and the jagged nature of the relationship between bad rate and k and D.

Fig 8.5.7: The effect of large D on the k-NN method.



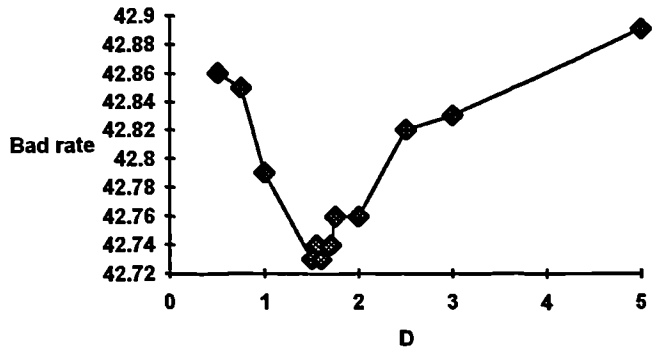


Figure 8.5.8: Graph of bad rate against D for metric d_6.

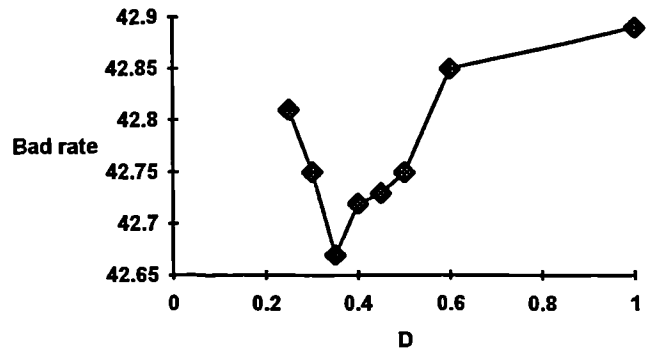


Figure 8.5.9: Graph of bad rate against D for metric d_7.

To make the last statement more evident and bring discussion of the raw data results together, Figures 8.5.10 - 8.5.13 show different plots of the bad rate against the range of k and D values together for metrics d_6 and d_7. This highlights the insensitivity of the minimum to the choice of both k and D.

Figure 8.5.10: Bad rate against k and D for metric d6

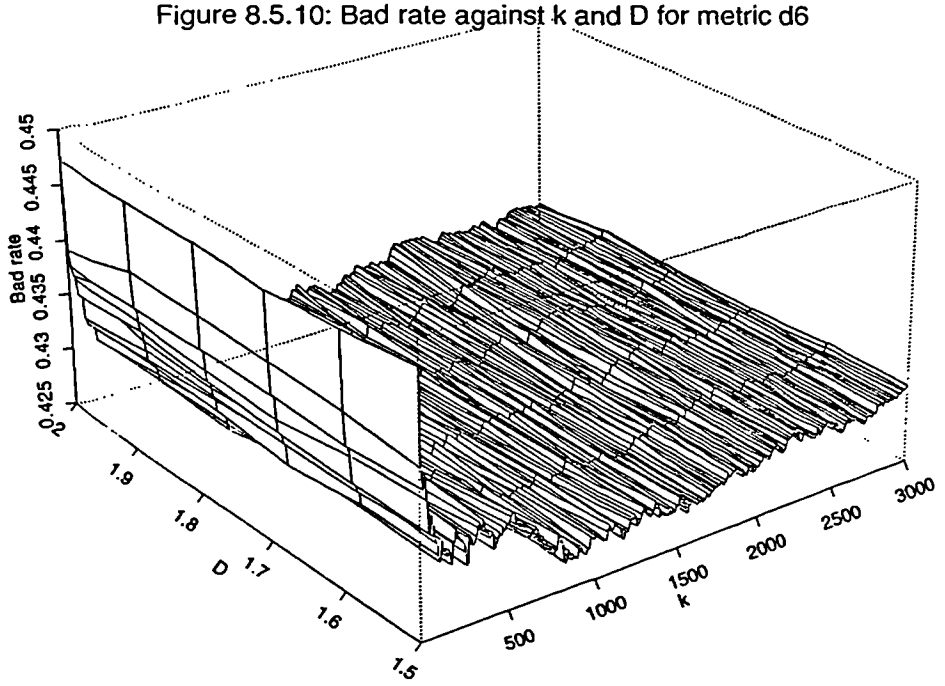


Figure 8.5.11: Bad rate against k and D for metric d6

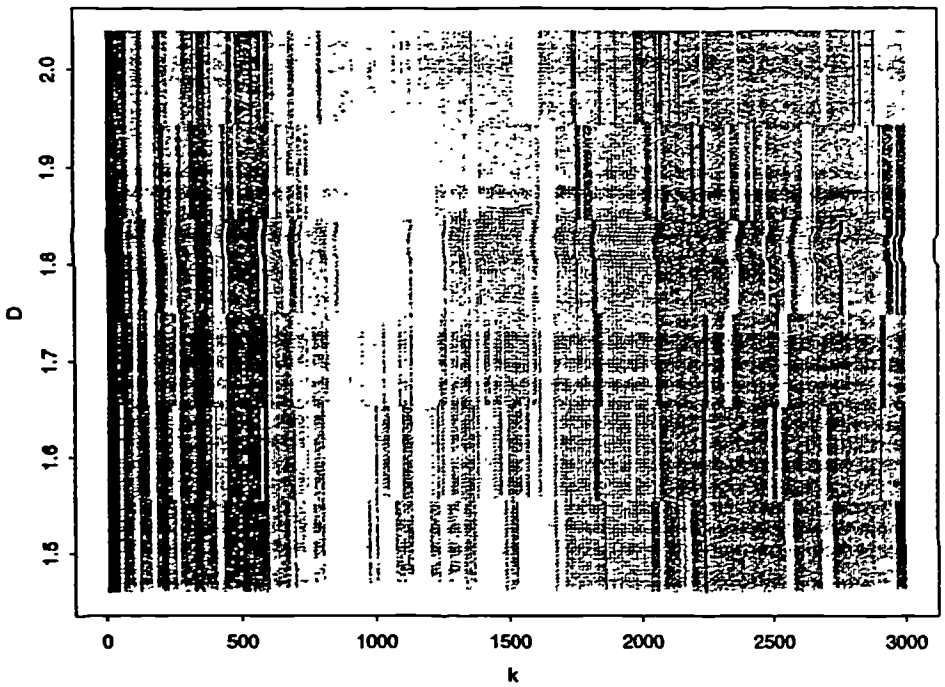


Figure 8.5.12: Bad rate against k and D for metric d7

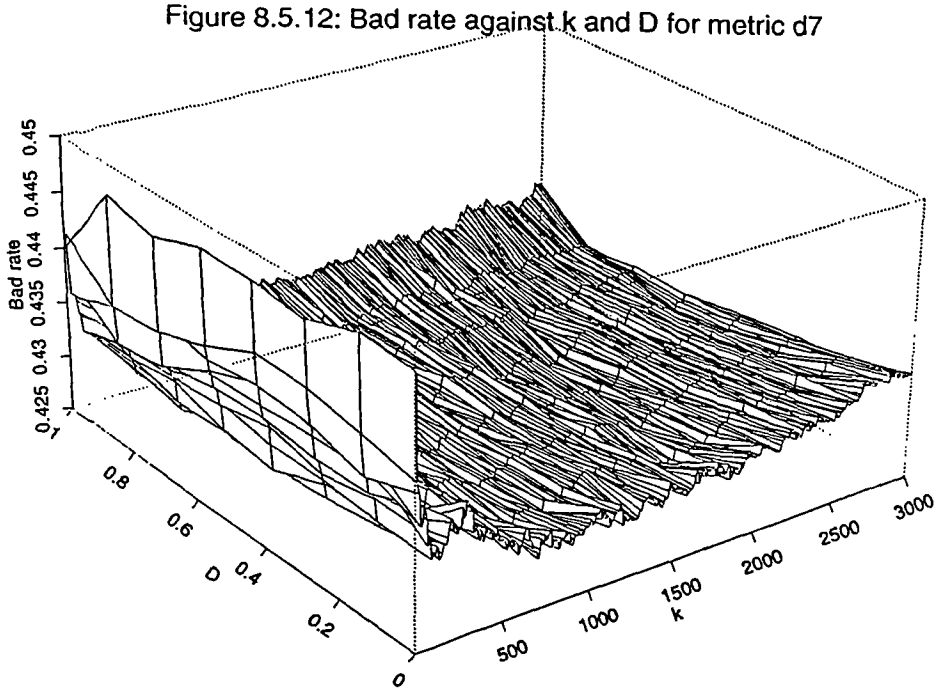
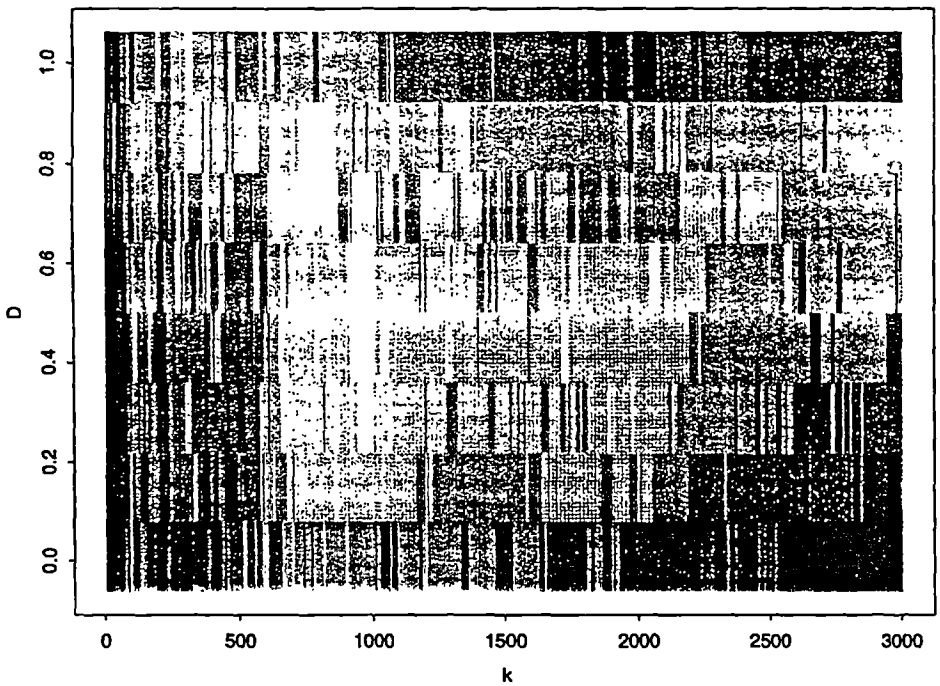


Figure 8.5.13: Bad rate against k and D for metric d7



8.5.2 *k*-NN results for the adjusted Euclidean metrics

Despite the insensitivity of the minimum bad rate to the choice of k and D, we choose to consider smoothed test set curves in order to reduce the jagged nature of the curve. By considering smoothed bad rate curves we hope to remove local minima that do not represent features of the population and thus obtain more robust estimates of the optimum parameters and performance of the k-NN method. This is important because, although the improvements from this approach are likely to be small (due to the insensitivity to k and D), a small reduction in bad debt can result in huge savings for the credit grantor.

To find the smoothed bad rate for a particular value of k we average the raw bad rates for a range of values of k around the value in question. The smoothing function is defined in Section 8.6.12. A parameter, h, is used to control the degree of smoothing. Our initial investigations showed that the results are fairly insensitive to the choice of h. In this study h was fixed as 20 because this value consistently gave near optimum performance using the criterion described above. As an example, Figures 8.5.14 and 8.5.15 show plots of smoothed bad rate against k for metrics d_6 and d_7 with the optimal D in each case. These plots suggest that we may be undersmoothing the bad rate curves. However, further investigations where k was selected from the design set showed that our choice of h is robust.

Tables 8.5.6 and 8.5.7 show the smoothed/unsmoothed bad rates and optimal k for metrics d_6 and d_7. We restrict attention to the values of D which give the best performance, together with results for the standard Euclidean metric (with $D = 0$). The bad rates from the smoothed curve are consistently higher than the minimum values from the unsmoothed curve as we would expect (since the raw data results represent an upper bound on performance). Our assertion is that the smoothed results give a more robust estimate of the performance that one can expect to achieve. However, we note that the optimum values of k for the raw and smoothed results are fairly similar (at about 800 - 1000). This indicates that the raw data results are giving a reasonable representation of the true performance of the k-NN rule. It remains to provide a practical method of selecting k and D (using the design set) and this is considered in Section 8.6.1.2.

Fig 8.5.14: Smoothed bad rate curve for metric d6 with D = 1.7

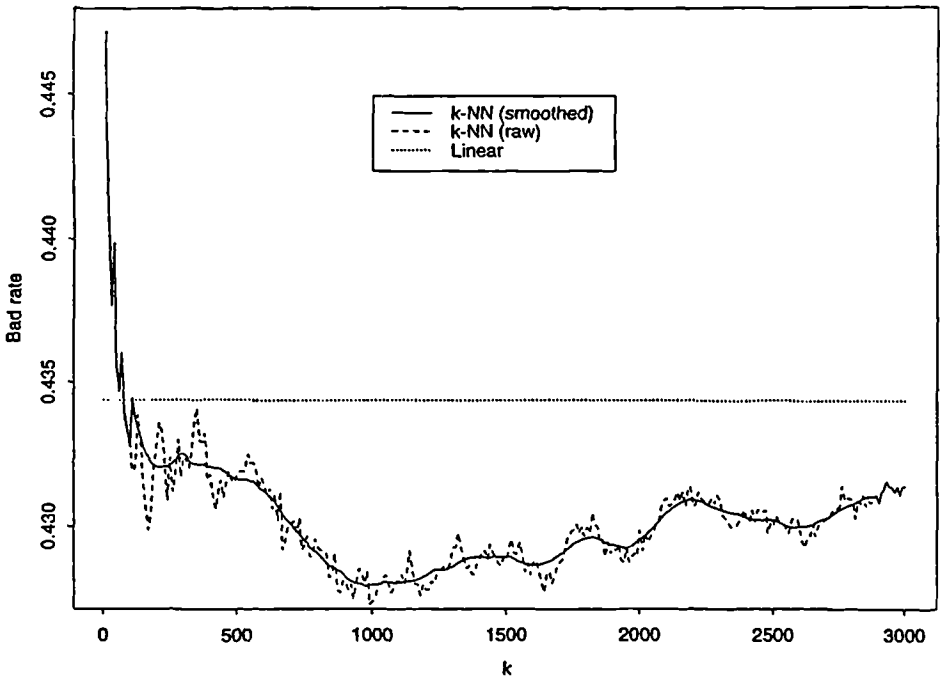


Fig 8.5.15: Smoothed bad rate curve for metric d7 with D = 0.4

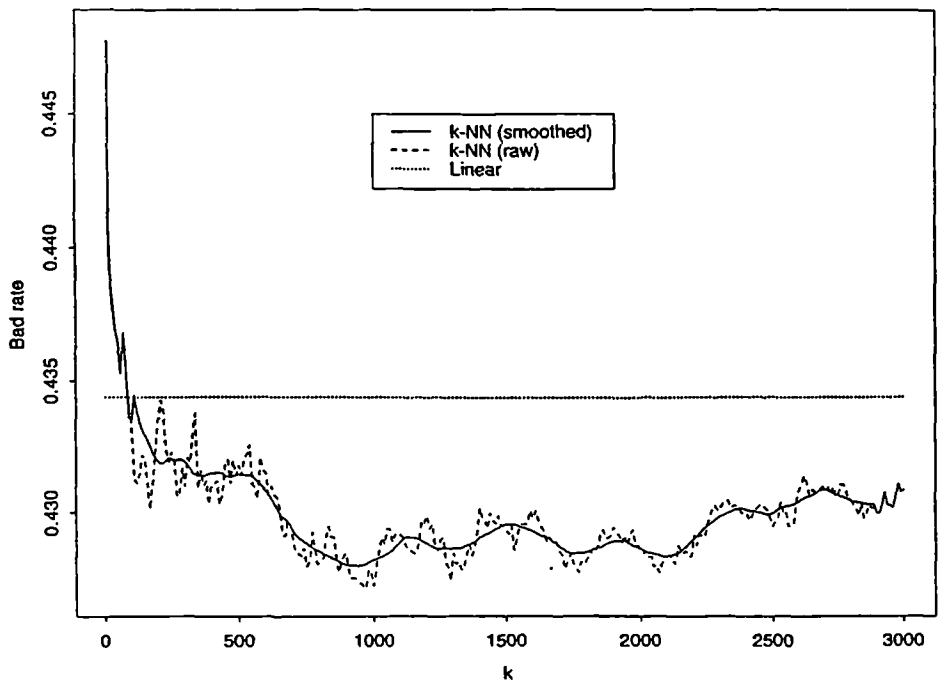


Fig 8.5.16: Smoothed bad rate curves for metric d6 with different D

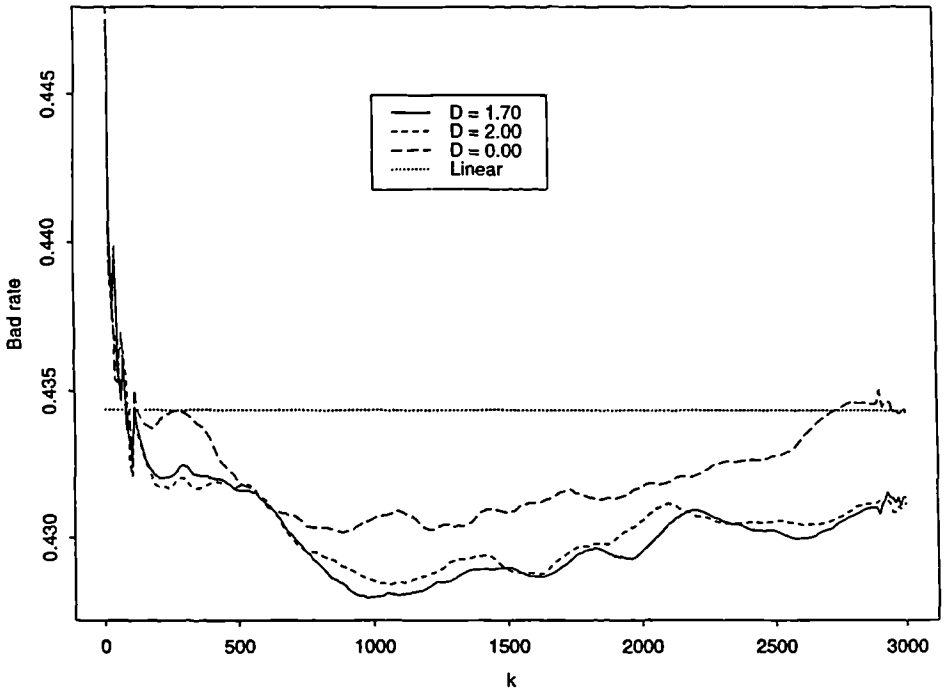


Fig 8.5.17: Smoothed bad rate curves for metric d7 with different D

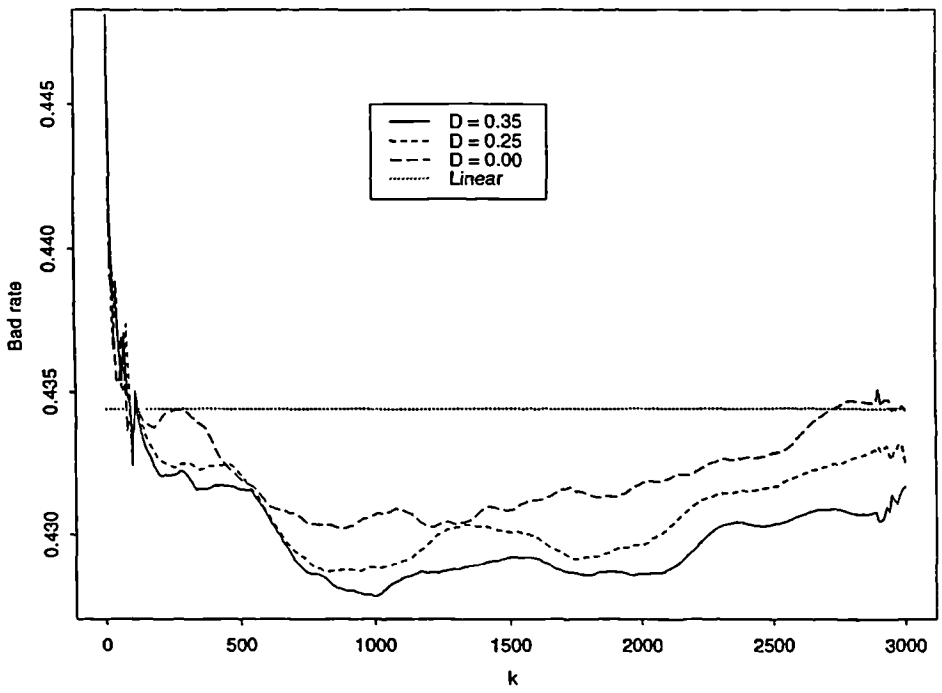


Figure 8.5.18: Smoothed bad rate against k and D for metric d6

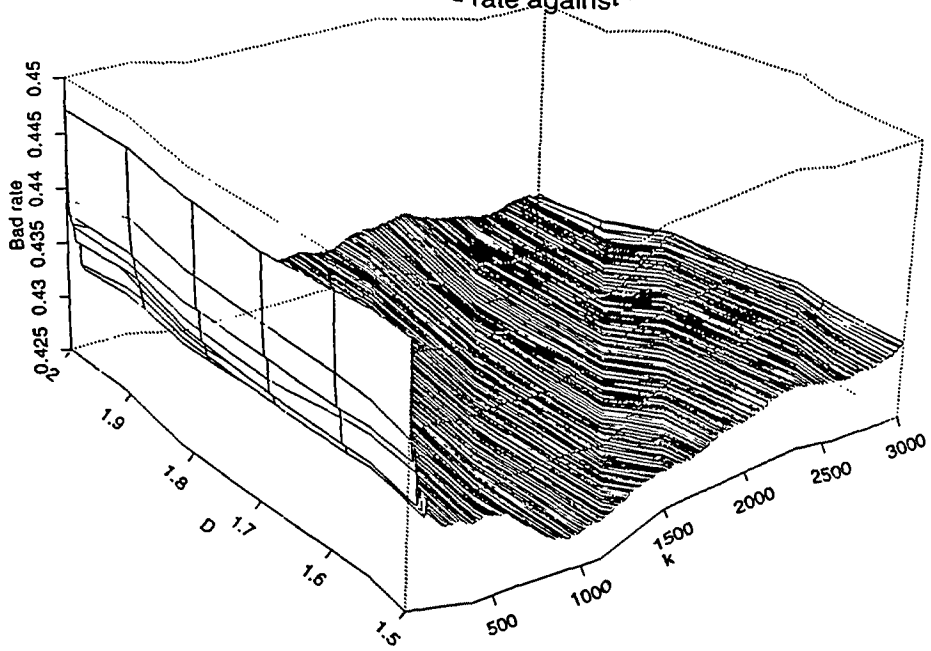


Figure 8.5.19: Smoothed bad rate against k and D for metric d7

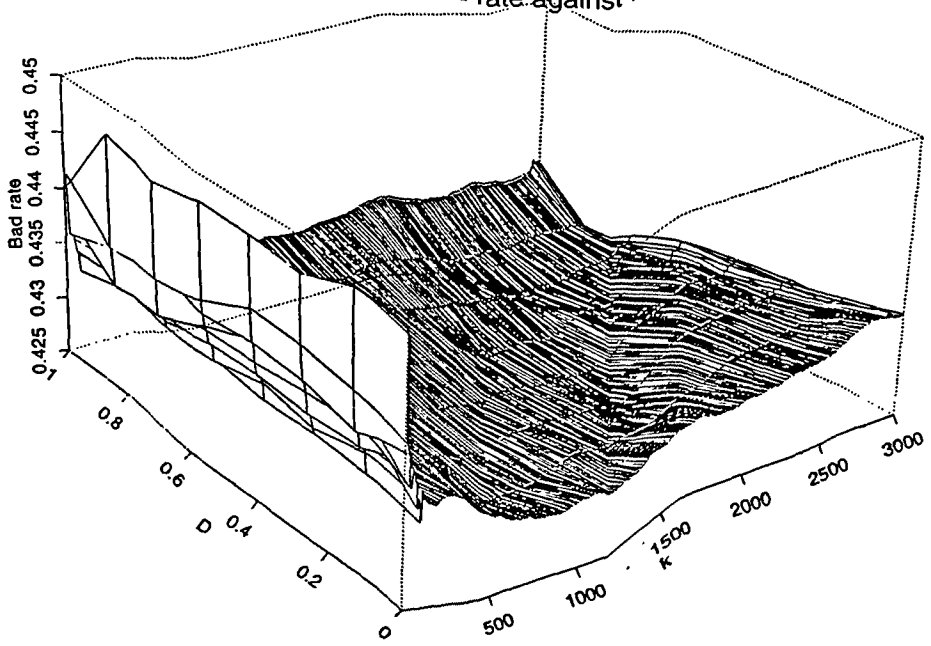
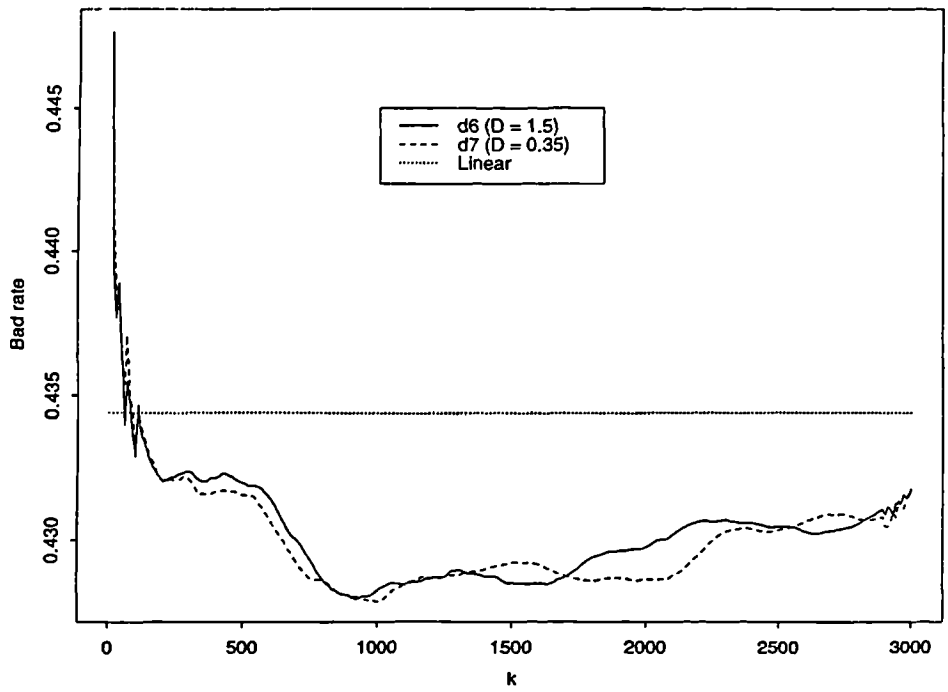


Fig 8.5.20: Smoothed bad rate curves for metrics d6 and d7 with optimal D



	Smoothed results		Unsmoothed results	
Value of D	Value of k	Bad rate	Value of k	Bad rate
0.00	820	43.03	700	42.93
1.50	920	42.79	830	42.73
1.55	920	42.81	930	42.73
1.60	980	42.81	1000	42.73
1.70	970	42.80	990	42.74

Table 8.5.6: Smoothed/unsmoothed bad rates for the adjusted Euclidean metric d_6.

	Smoothed results		Unsmoothed results	
Value of D	Value of k	Bad rate	Value of k	Bad rate
0.30	960	42.81	870	42.75
0.35	1000	42.78	930	42.67
0.45	910	42.85	970	42.72
0.50	760	42.87	820	42.75

Table 8.5.7: Smoothed/unsmoothed bad rates for the adjusted Euclidean metric d_7.

The tables show that the optimal bad rates from the smoothed k-NN results are below the linear regression result (43.44%). This indicates that, after taking into account peculiarities of the test sample, there is a range of values of k and D for which our k-NN method can outperform regression. It just remains to identify these ranges of D and k values from the design set so that the method can be applied in practice. This provides the motivation for a more extensive comparison study in Section 8.6.

The results also indicate that the adjusted Euclidean metrics d_6 and d_7 can give improved performance over the standard Euclidean metric ($D = 0$). The lowest smoothed bad rate for the adjusted Euclidean metric is 42.81 compared to 43.03 for the standard Euclidean metric.

Figures 8.5.16 and 8.5.17 show plots of smoothed bad rate against k for a range of D values for the two metrics. We note that, in both cases, the curve for the optimal D value (1.7 for d_6 and 0.35 for d_7) is below the other curves

for almost all values of k. This consistency gives us grounds for believing that we have smoothed out peculiarities of the data and are describing properties of the underlying population. It is also further evidence that an optimal value of D exists.

Figures 8.5.18 and 8.5.19 show smoothed bad rate against the range of k and D values for the two metrics. They are analogous to the raw data plots shown in Figures 8.5.10 and 8.5.12. The insensitivity of the minimum to the choice of k and D is now even more apparent. However, metric d_6 does appear to have a slightly flatter minimum than metric d_7. This may be partly due to the different scaling of the parameter D under the two metrics.

We end by noting that the metrics d_6 and d_7 give very similar performance (but the optimal values of D are different - see Tables 8.5.6 and 8.5.7). Figure 8.5.20 shows the smoothed bad rate curves for the optimal values of D under the two metrics. The curves look very similar in their optimum regions (circa $k = 1000$) and give almost identical bounds on performance. We choose to use the metric d_6 in the rest of this chapter because of its stronger theoretical basis.

8.5.3 Other metrics

Until this point we have concentrated exclusively on the adjusted Euclidean metrics. We have seen that by giving extra emphasis to distance orthogonal to equi-probability contours, it is possible to improve upon the performance of the standard Euclidean metric. In Section 8.2.3.5 we discussed work by Todeschini (1989) which indicated that the city block metric (amongst others) can lead to improved performance over the standard Euclidean metric. For this reason we investigate the performance of the city block metric and the more general Minkowski distance. We also consider data dependent versions of these metrics by applying the transformation of the data defined in Section 8.3.2.

Figure 8.5.21 shows four bad rate curves for the city block metric d_5 as described below:

(1) The city block metric is shown with $D = 0$.

Fig 8.5.21: City-block metric (raw data)

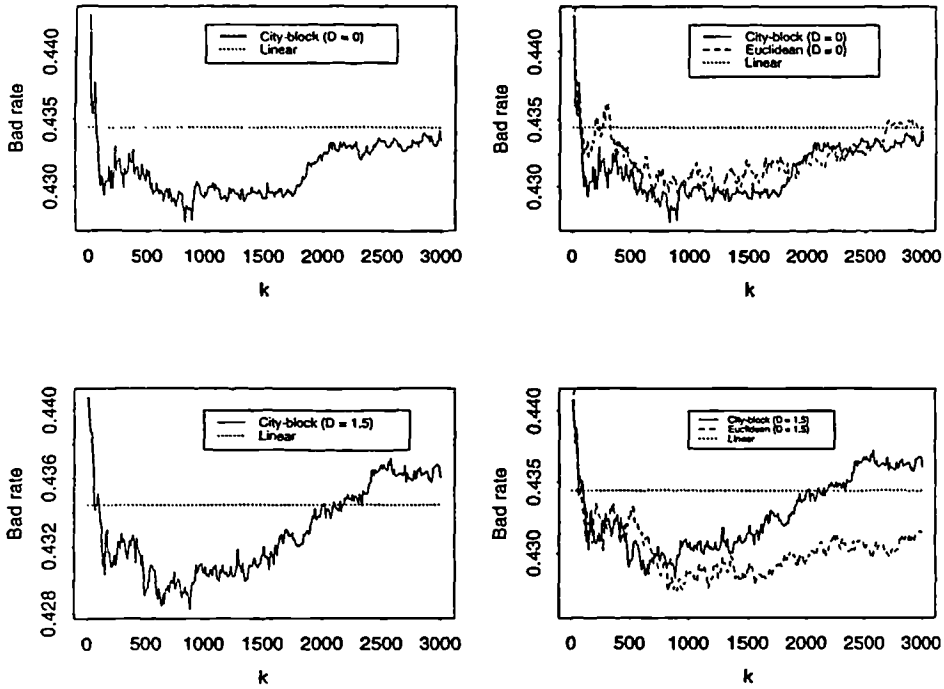
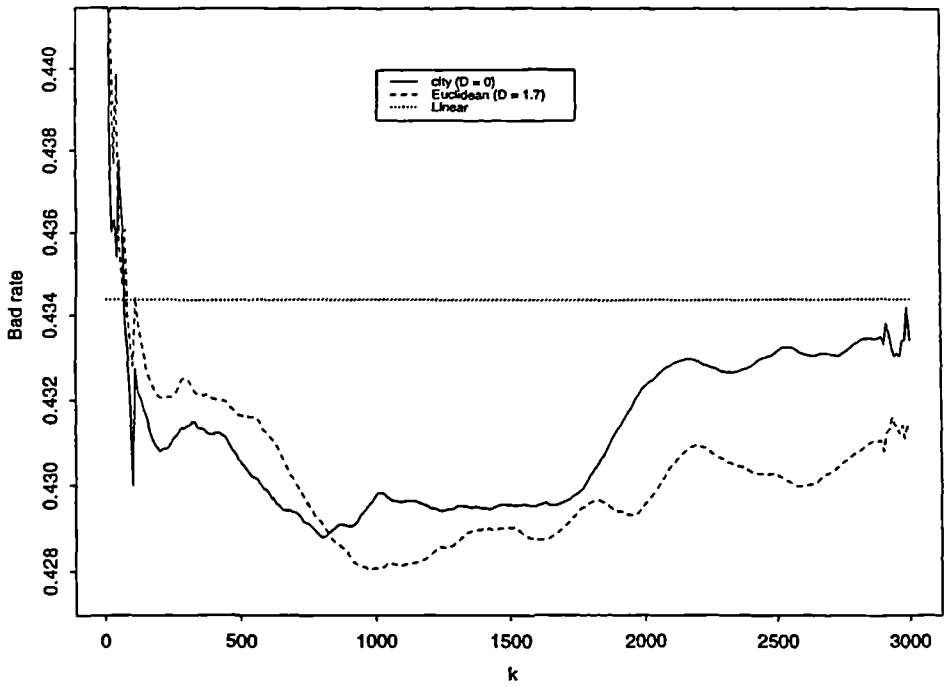


Fig 8.5.22: City-block metric v Euclidean metric (using optimal D)



(2) The curve for the Euclidean metric with $D = 0$ is superimposed. For this value of D, the city block metric is superior to the Euclidean metric for most values of k.

(3) The city block measure is shown with $D = 1.5$ (a value in the optimal region).

(4) The curve for the Euclidean metric with $D = 1.5$ is superimposed. For this value of D, and other values in the optimal region, the Eucliudean metric outperforms the city block metric for most k.

The above plots show that if we use the standard Euclidean and city block metrics (with $D = 0$) then the city block metric gives the best performance. However, if we use the data dependent versions of the two metrics (using the same value of D in each case) then the adjusted Euclidean metric gives the best performance.

To make a more useful comparison of the two metrics, we selected the optimal values of D for each metric and then compared the results. Figure 8.5.22 shows the resulting smoothed plot for the test set.

The adjusted Euclidean metric gives better performance over the region of optimal k (although the city block performs better for lower k). Table 8.5.8 summarises the bounds on performance from the smoothed and unsmoothed curves.

		Smoothed results		Unsmoothed results	
	Value of D	Value of k	Bad rate	Value of k	Bad rate
Euclidean	1.70	1060	42.81	990	42.74
City block	0.00	800	42.88	830	42.74

Table 8.5.8: Bounds on performance for the k-NN method using the adjusted Euclidean and city block metrics with optimal D.

The table shows that, although the bounds are identical for the unsmoothed results, the smoothed bound is slightly lower for the adjusted Euclidean metric. We also note that the adjusted Euclidean metric has a flatter minimum than the city block metric (see Figures 8.5.21 and 8.5.22). For these reasons we believe that by using an adjusted version of the Euclidean metric we have been able to improve performance over both the standard Euclidean metric and the city block metric.

Further work was carried out to investigate the influence of the parameter r in the general Minkowski distance d_4. To start with we restricted attention to the case where $D = 0$ to reduce the number of variables in the analysis. Figure 8.5.23 shows the raw bad rate curves for four values of r.

There is noticeable variation in the shape of the curves for different r. In order to assess whether these are real differences, we smoothed the bad rate curves to remove sample variation. Figure 8.5.24 shows the smoothed bad rate against k and r. This indicates the insensitivity of bad rate to both k and r.

Table 8.5.9 shows the bounds on performance for different r.

Value of r	Smoothed results		Unsmoothed results	
	Lowest bad rate	Value of k	Lowest bad rate	Value of k
0.4	42.80	1320	42.71	1370
0.5	42.83	1200	42.71	1250
0.6	42.85	1260	42.79	1010
0.8	42.92	890	42.82	470
1	42.88	800	42.74	830

Table 8.5.9: Bad rates for the general Minkowski distance d_4 for different r.

The results show that the Minkowski distance metrics with $r = 0.4$ and 0.5 give comparable bounds to the adjusted Euclidean metrics d_6 and d_7.

We also applied the transformation of the data defined in Section 8.3.2 to the general Minkowski distance measure to see if this could lead to improved

Fig 8.5.23: The general Minkowski distance (for different r).

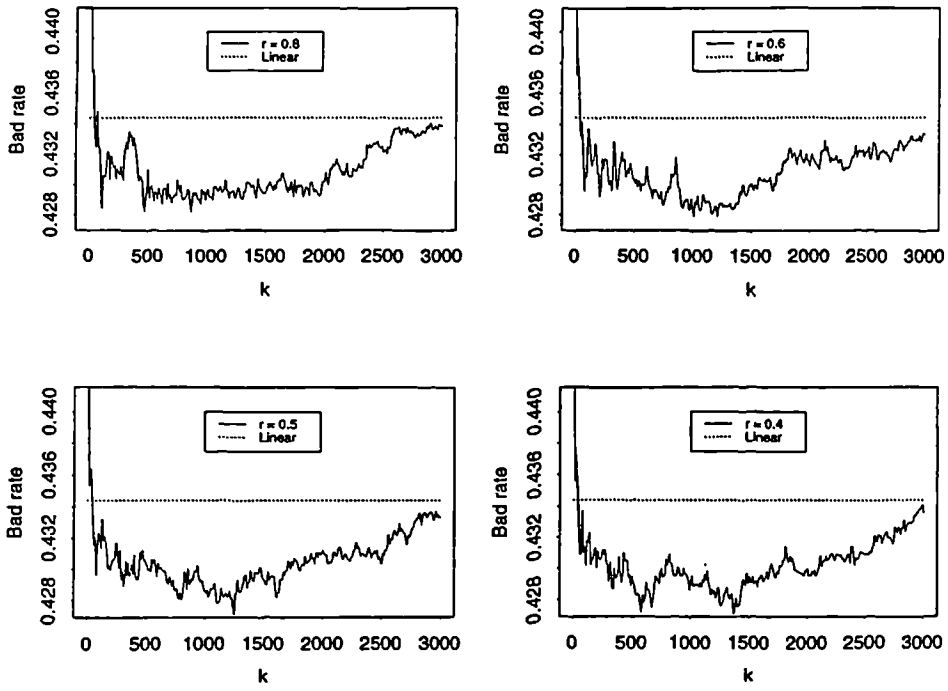
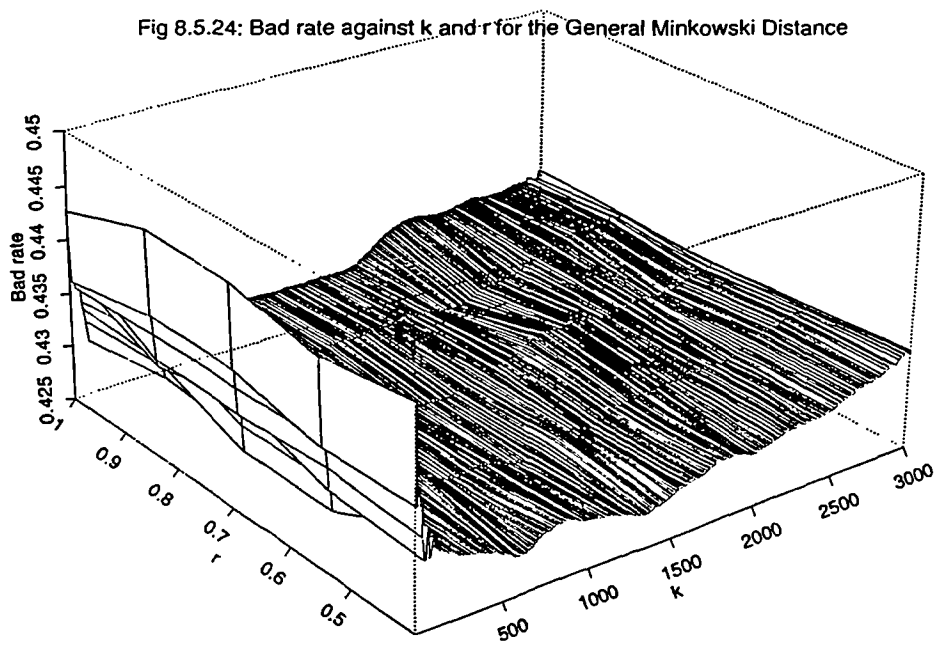


Fig 8.5.24: Bad rate against k and r for the General Minkowski Distance



discrimination. We found that, as with the adjusted Euclidean metric, the adjusted general Minkowski distance is insensitive to the choice of D. A value of 1.6 was found to give near optimal results across the range of r values between 0 and 1. Table 8.5.10 shows the resulting "best" bad rates amongst the accepts in the test set for different values of r.

Value of r	Smoothed results		Unsmoothed results	
	Lowest bad rate	Value of k	Lowest bad rate	Value of k
0.1	42.95	370	42.81	280
0.2	42.75	430	42.64	440
0.3	42.75	480	42.65	480
0.4	42.73	560	42.63	520
0.5	42.77	390	42.70	360
0.6	42.71	390	42.57	360
0.8	42.80	720	42.66	660
1	42.94	580	42.84	530

Table 8.5.10: Bad rates for the general Minkowski distance d_4 for different r given $D = 1.6$.

The table shows that the transformation of the data has led to improved performance of the general Minkowski distance metric for all the values of r considered apart from $r=1$. Furthermore, the metric gives better optimal performance than the adjusted Euclidean metric for $0.2 \leq r \leq 0.8$, although the differences are small. In the rest of this chapter we focus on using the adjusted Euclidean metric because the resulting k-NN performance is less sensitive to k. Further work is needed on the application of the k-NN method with the general Minkowski distance metric to credit scoring and, in particular, on the selection of optimal values of the parameter r.

8.5.4 Explanations for the high optimal k

One of the features of the k-NN results that holds for all the metrics considered is the high optimal value of k. In this section we try to explain this by examining the estimates of $P(g|\mathbf{x})$ using a graphical approach.

As we discussed in Section 8.2.4 the choice of a suitable value of k involves a trade-off between minimising the variance and bias of the estimates of $P(g|\mathbf{x})$. The variance of $\hat{P}(g|\mathbf{x})$ is given by

$$\text{var}(\hat{p}) = p(1-p)/k$$

where $p = P(g|\mathbf{x})$. Therefore, as k increases, so the variance decreases. On the other hand, the bias of the estimate may increase as k increases, as points more distant from $\mathbf{x}$ are considered.

We consider two possible explanations for the high optimal value of k both of which relate to the bias of the estimates $\hat{P}(g|\mathbf{x})$:

(1) The bias of the probability estimates may be unimportant as long as the order of $\hat{P}(g|\mathbf{x})$ is the same as the order of $P(g|\mathbf{x})$.

(2) The structure of the population is such that $P(g|\mathbf{x})$ lies between 0.2 and 0.8 for a higher than expected region of the characteristic space. This implies that bias might be expected to increase relatively more slowly than variance decreases, resulting in a high optimal k.

For the second of these explanations to hold, we need to confirm the slope of $P(g|\mathbf{x})$ and show that the bias of the good/bad probability estimates is relatively small for the optimal k (approx 1000). For the first explanation to hold the bias can be quite large for the optimal k.

8.5.4.1 Bias of the $P(g|\mathbf{x})$ estimates

We now take a graphical approach to estimating the bias of the predicted probabilities from the k-NN method. We started by reducing the data to one dimension by fitting a linear regression model. Figure 8.5.25 shows four plots designed to compare the true $P(g|\mathbf{x})$ with the estimated probabilities from the k-NN rule for k in the optimal region (fixed at 1000 in this study):

(1) An estimate of the true $P(g|\mathbf{x})$ was obtained by taking kernel smoothed estimates of the sample responses at each regression score. A key assumption is that this gives an accurate representation of the true good/bad probabilities at each regression score.

Fig 8.5.25: A graphical comparison of the accuracy of the estimated P(g|x) from the *k*-NN method and logistic regression.

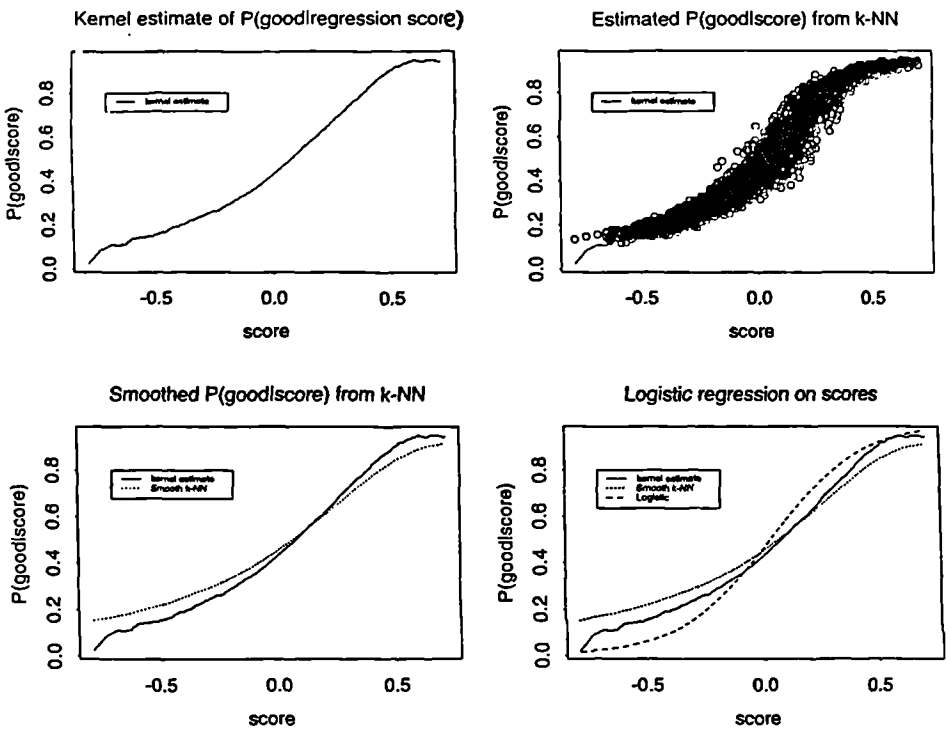


Fig 8.5.26: Smoothed P(goodlscore) from k-NN with k > 1000

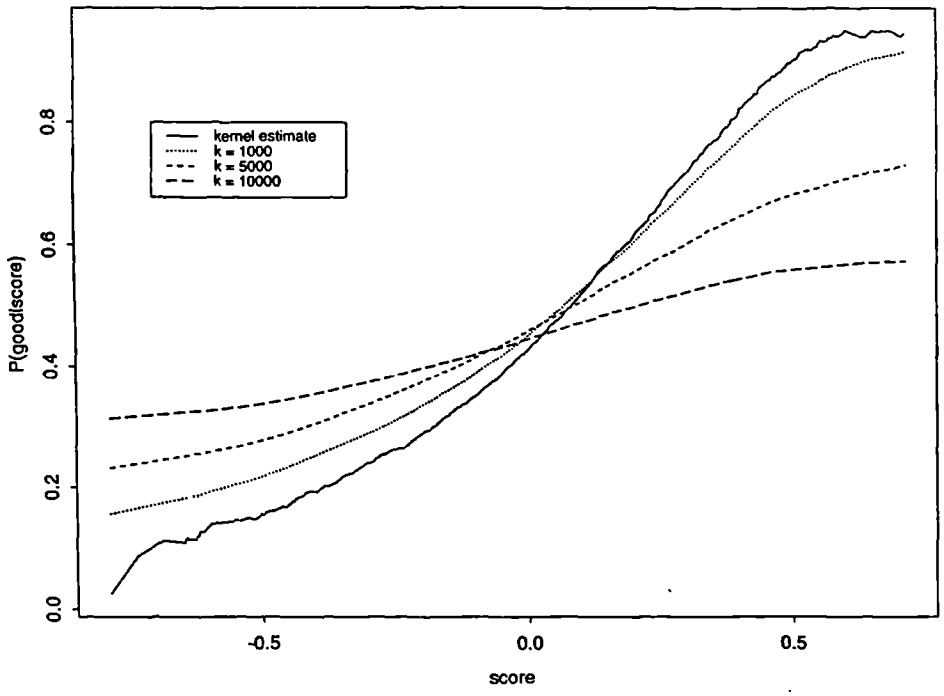


Fig 8.5.27: Smoothed P(goodIscore) from k-NN with k < 1000

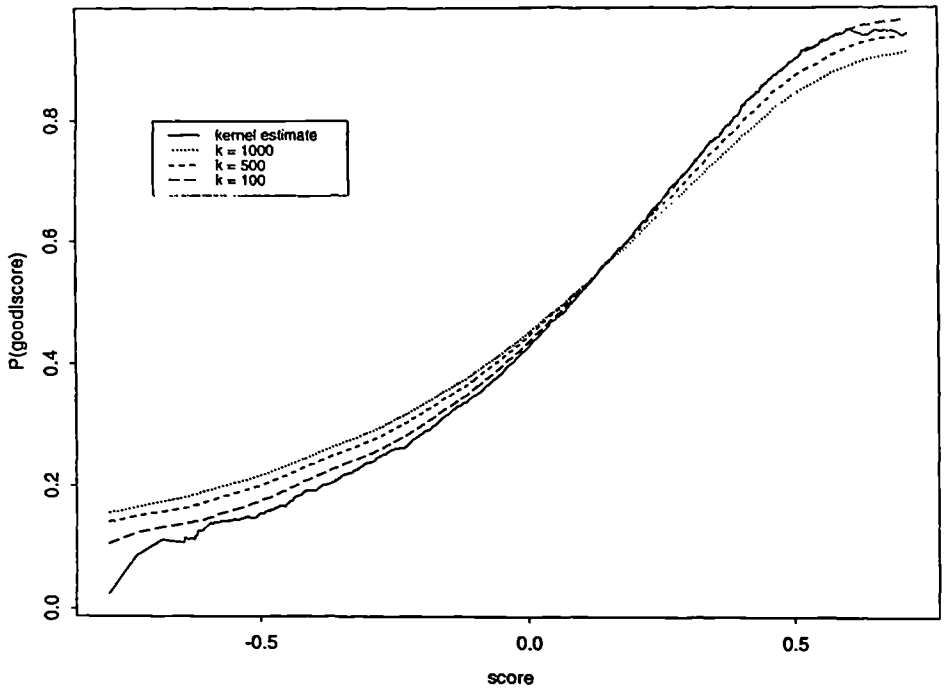


Fig 8.5.28: Estimated bias for different k values

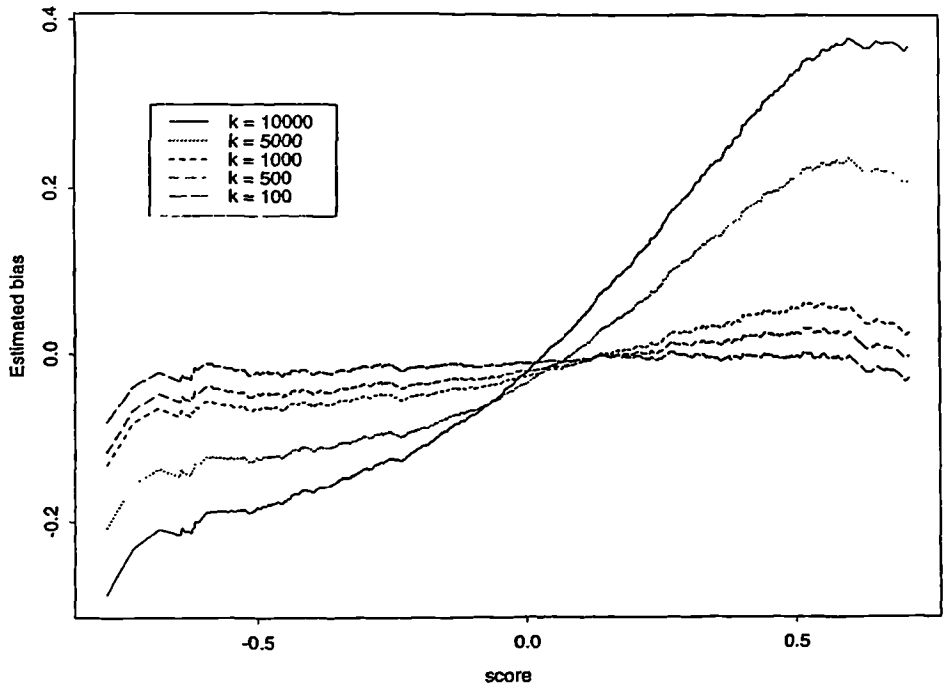


Fig 8.5.29: Histogram of scores with estimated P(goodIscore)

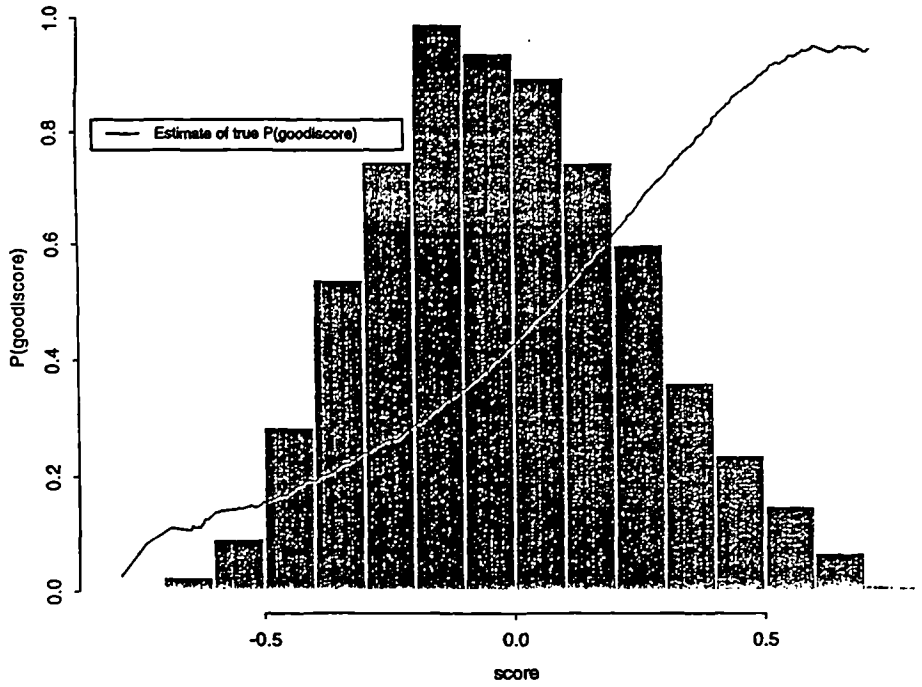


Fig 8.5.30: Histogram of logistic scores with estimated P(goodIscore)

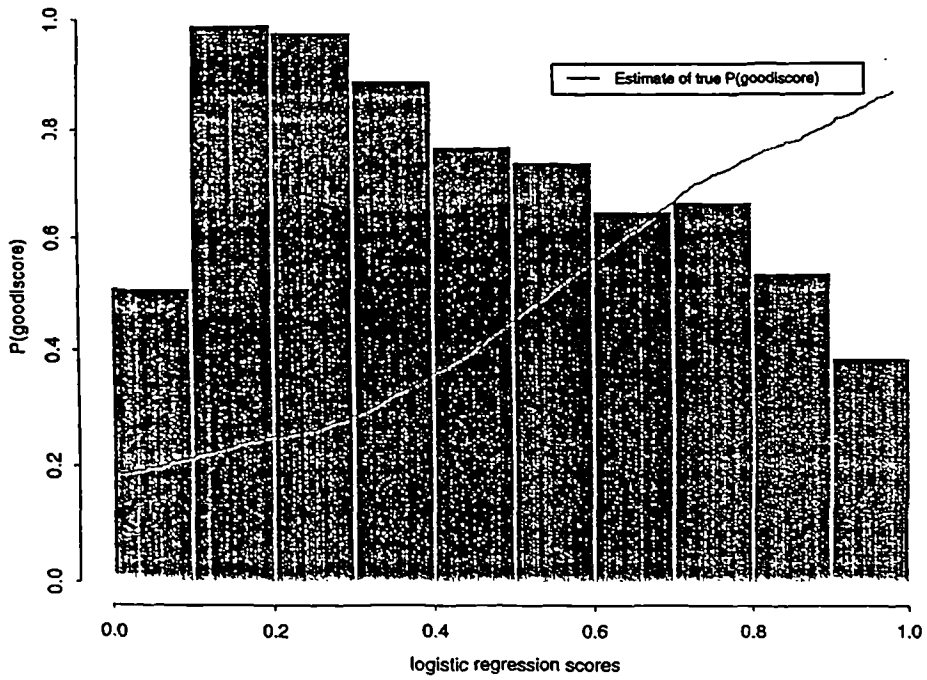
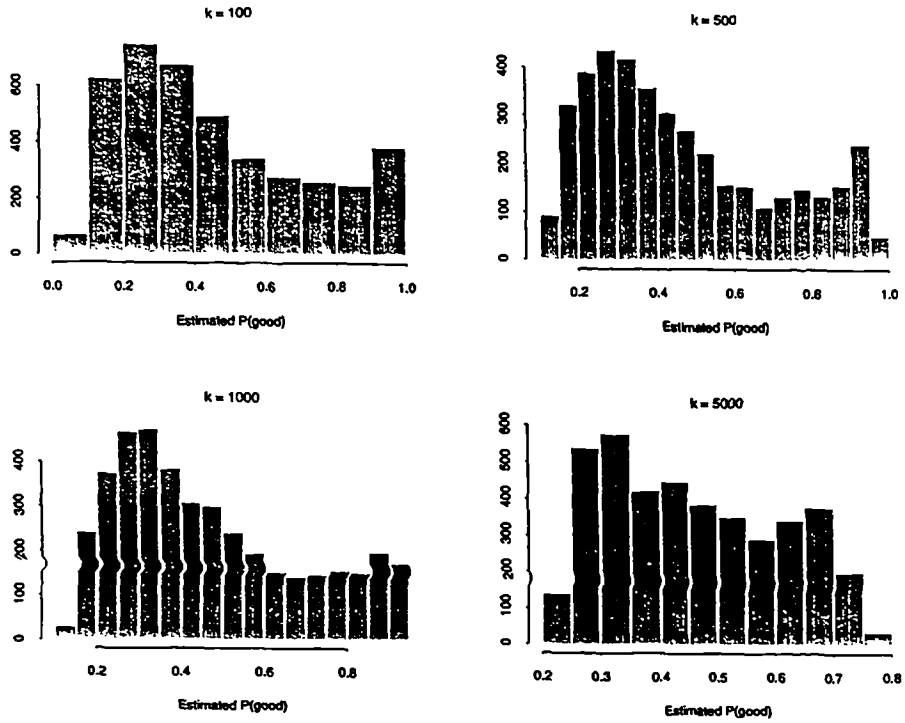


Fig 8.5.31: Histograms of P(g|x) for the k-NN method with different k.



(2) Estimates of $P(g|\mathbf{x})$ from the k-NN method (using the full characteristic space) are superimposed for each member of the test sample. The probability cloud represents variation in the creditworthiness of points not accounted for by the linear regression scores. We note that there is more variation in the probability estimates for mid-range scores. In order to provide a more robust classification rule, we could vary k across the characteristic space according to the variance of the probability estimates.

(3) The variation described in (2) is removed by taking a kernel smoothed estimate of the k-NN probability estimates at each regression score.

(4) A logistic regression model was fitted to the linear regression scores using the true classes and this is superimposed. It shows greater bias from the "true" $P(g|\mathbf{x})$ than the k-NN estimates do.

The above figure shows that the k-NN probability estimates for optimal k have lower bias than the logistic regression estimates. To further explore how the bias of the k-NN rule changes with k, we plotted estimated $P(g|\mathbf{x})$ for different values of k. Figure 8.5.26 shows the k-NN curves for k values greater than the optimum. Figure 8.5.27 shows k-NN curves for k less than the optimum. The kernel estimate of the true $P(g|\mathbf{x})$ is added to both plots. Figure 8.5.28 shows the estimated bias for the range of k values (i.e. difference between kernel estimate and k-NN estimate)

The plots show that bias does indeed increase as k increases. The optimum k does give more biased estimates than the sort of values advocated by Enas and Choi (1986). However, the magnitude of the bias for $k = 1000$ is not sufficient to reject our second explanation for the high optimal k.

8.5.4.2 Slope of P(g|x)

Figure 8.5.29 shows a histogram of the linear regression scores with the kernel estimate of the true $P(g|\mathbf{x})$ superimposed. 14% of the test sample lies outside the interval $0.2 < P(g|\mathbf{score}) < 0.8$. If the data were uniformly distributed across [0,1] then we would expect 40% of the sample to lie outside this

interval. This indicates that the slope of $P(g|\mathbf{x})$ is shallow, thus satisfying the condition required for our second explanation from above.

To confirm this condition, we repeated the procedure outlined above using logistic regression scores instead of linear regression scores as a method of dimension reduction. Figure 8.5.30 shows the histogram of scores with the kernel estimate of the of the true $P(g|\mathbf{x})$ superimposed. In this case, only 9.27% of the sample lies outside the interval $0.2 < P(g|\mathbf{score}) < 0.8$, giving a further indication of the shallow nature of $P(g|\mathbf{x})$.

The preceding analysis has provided some evidence that the slope of $P(g|\mathbf{x})$ is shallow, with the probabilities in the range [0.2,0.8] for most of the characteristic space. This is combined with evidence from Section 8.5.4.1 of some bias at the optimum k. The bias was less than the bias of the probability estimates from the logistic regression. These factors combine to suggest that our second explanation for the high optimal k is most likely.

To complete our discussion of the nature of the predicted probabilities of class membership from the k-NN method, we looked at histograms of predicted probability for different values of k. These are shown for k = 100, 500, 1000 and 5000 in Figure 8.5.31. We observe that as k increases so the estimated probabilities occur on a smaller range (this corresponds to the increase in bias). The histograms are bimodal and the largest mode occurs for low proportion good. It is surprising that there are two modes because we believe that the population of credit applicants cannot be sub-divided into distinct sub-populations and that instead applicants can be arranged somewhere on a good-bad continuum. This phenomenon may be due to the omission of the class of "other" applicants from the analysis.

8.5.5 Standard credit scoring techniques

In the earlier parts of this section we have used linear regression to provide a baseline against which to assess the k-NN method with adjusted Euclidean metrics. To provide a more representative picture of the performance of standard credit scoring techniques, we now consider the results of applying logistic regression, decision trees and decision graphs. For a discussion of the

properties of these techniques see Section 4.9 and Chapter 7. Table 8.5.11 shows the bad rates at a 70% acceptance rate. Results are shown for the decision tree method using look-ahead 1 and 2 and for the decision graph method using different values of the join parameter.

Classification method	Bad rate
Linear regression	43.44
Logistic regression	43.37
Decision tree (LA1)	44.50
Decision tree (LA2)	43.88
Decision graph (P_j =0.05)	43.88
Decision graph (P_j =0.1)	43.43
Decision graph (P_j =0.2)	43.78

Table 8.5.11: Bad rates at a 70% acceptance rate for linear regression, logistic regression and various decision trees/graphs.

Logistic regression gives the lowest bad rate but only by a very small margin. To confirm that there is no significant difference between the linear and logistic regression results we performed the two significance tests of Chapter 5. The swapsets are given by Table 8.5.12.

	bads	goods
Linear	16	6
Logistic	18	4

Table 8.5.12: 2*2 classification table for applicants accepted by either linear or logistic scorecard (but not both).

Neither test showed a significant difference between the bad rates. We also note that the best decision graph gives almost identical performance to the linear regression model.

We conclude that the range of standard classification techniques considered give very similar performance for this data set and the criterion used. We have seen in Section 5.3 that the k-NN method has the potential to provide improved discrimination: the minimum bound on performance was 42.67 for metric d_7 with $D = 0.35$. In Section 6 we use repeated sampling to test whether it is

possible to provide a practical k-NN classification rule that can approach the bounds on performance.

8.6 An empirical study of the application of the k-NN method to credit scoring

There are two limitations of the analysis in Section 8.5:

- We did not pre-select the parameters k and D from the design set before applying the classifier to the test set.

- The results are subject to sample variation.

We now present a further study of the k-NN method with adjusted Euclidean metrics designed to satisfy these two considerations. In order to reduce the influence of sample variation our analysis involves taking 5 random design/test sample splits from the combined design and test samples described in Table 8.5.1 (using the same ratio between sample sizes). The method is then applied to each of the design sets in turn and the results are assessed using the corresponding test set. If the results are then averaged our assessment procedure becomes similar to that adopted by Leonard (1993). This approach is often called (n-fold) cross-validation.

8.6.1 Estimation of k and D

In order to apply our k-NN methodology to future samples it is necessary to choose values for the two parameters k and D. We propose a procedure for doing this using the design set and assess performance using an independent test sample from the same population.

In principle the estimation of suitable k and D is straightforward, requiring a simple bivariate optimisation (for example using a steepest descent algorithm) of the criterion. In what follows we present a more detailed discussion of the properties of this criterion for different choices of k and D so that better understanding is gained.

8.6.1.1 Bad rate curves

Figures 8.6.1 to 8.6.5 show plots of bad rate against k for the 5 randomly selected design/test sets using various choices of D. The figures were chosen to illustrate the subtle variations in the bad rate curves across different samples and values of D. For each curve the bad rate amongst the accepts is shown using both the design and test samples. The design sample curve comes from classifying each design set point using its k nearest design set points (excluding itself). The design set curve could be used to select the appropriate value of k to consider for the corresponding test set. The resulting k are shown by the vertical lines joining the lowest points on the design set curves with the appropriate test curves (we consider alternative ways of selecting k after a discussion of properties of these curves). The horizontal broken lines represent the performance of a classifier using linear regression.

Figure 8.6.6 shows the averaged bad rate curve for the five samples, using, in each case, the optimal value of D as estimated from the design set.

The observations made in Section 8.5.1 about the jagged nature of the bad rate curves, the high optimal k and the breadth of the minimums apply equally to these curves. We add a couple of points:

(i) There is a difference in level between the bad rate curves for the design and test samples in Figures 8.6.1 to 8.6.6. This is explained by the difference in bad rates amongst the full design and test samples in these cases. For example, sample 1 has a bad rate of 54.24% in the design sample and 55.59% in the test sample. We chose not to fix the proportion of goods in each randomly selected test sample because we were interested in prediction performance on future applicants and the proportion of bads in such a sample will vary.

(ii) We found that for samples 2-5 there exist values of D and ranges of k for which the test sample curves drop below the line representing the performance of a classifier using linear regression. This indicates that our method has the potential to lead to improved discrimination.

Fig 8.6.1: k-NN using data set 1 and D = 0.00

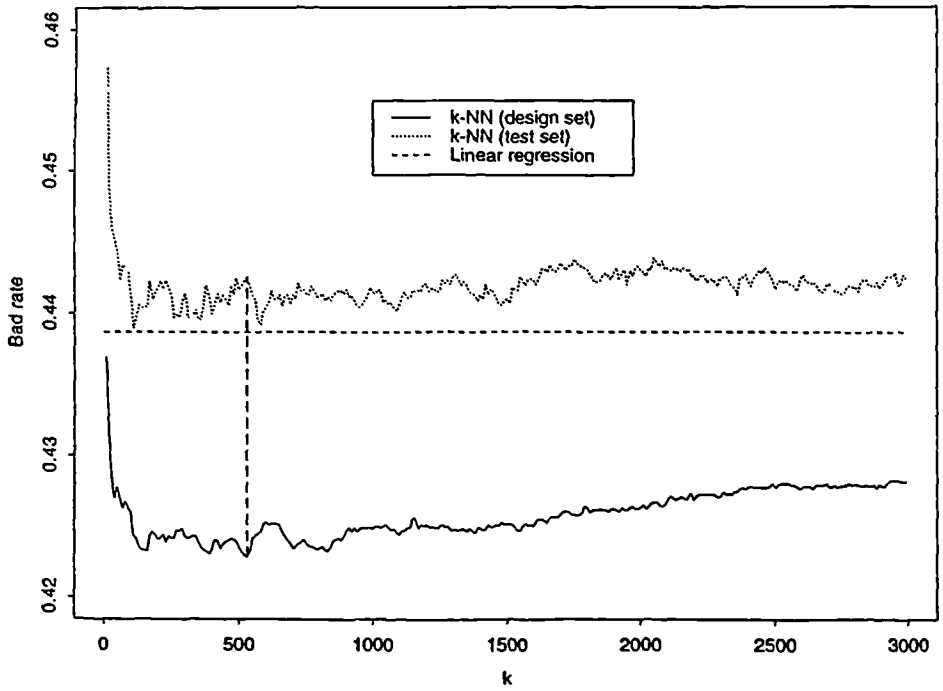


Fig 8.6.2: k-NN using data set 2 and D = 1.60

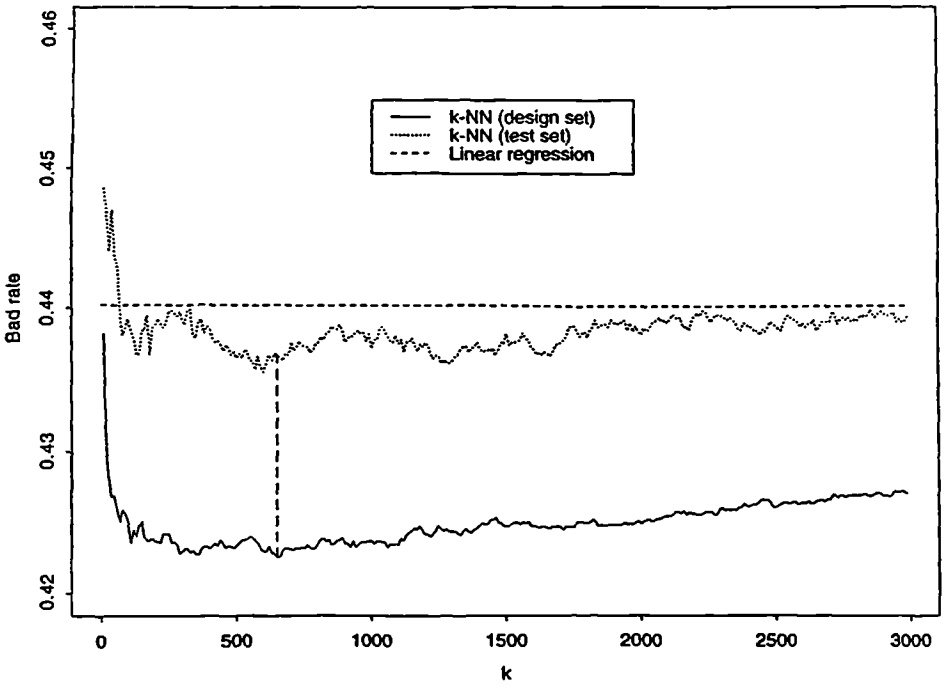


Fig 8.6.3: k-NN using data set 3 and D = 1.40

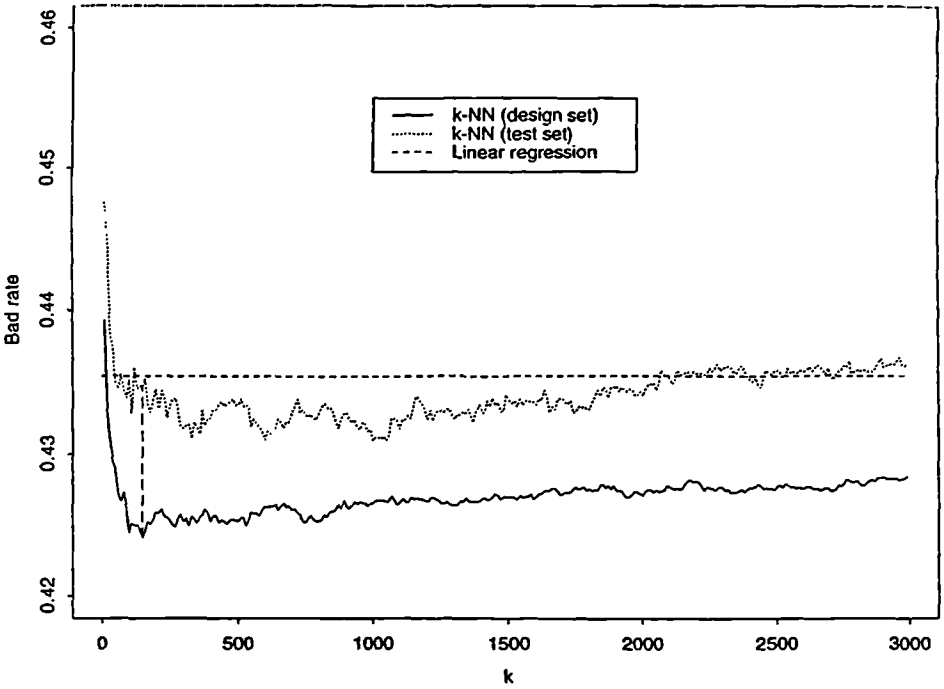


Fig 8.6.4: k-NN using data set 4 and D = 1.80

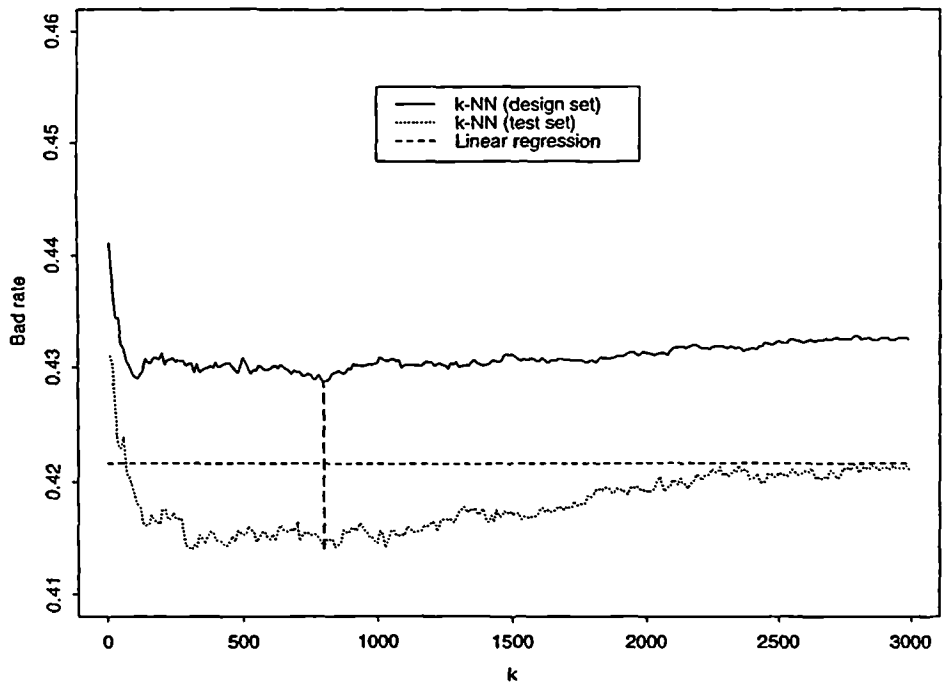


Fig 8.6.5: k-NN using data set 5 and D = 1.60

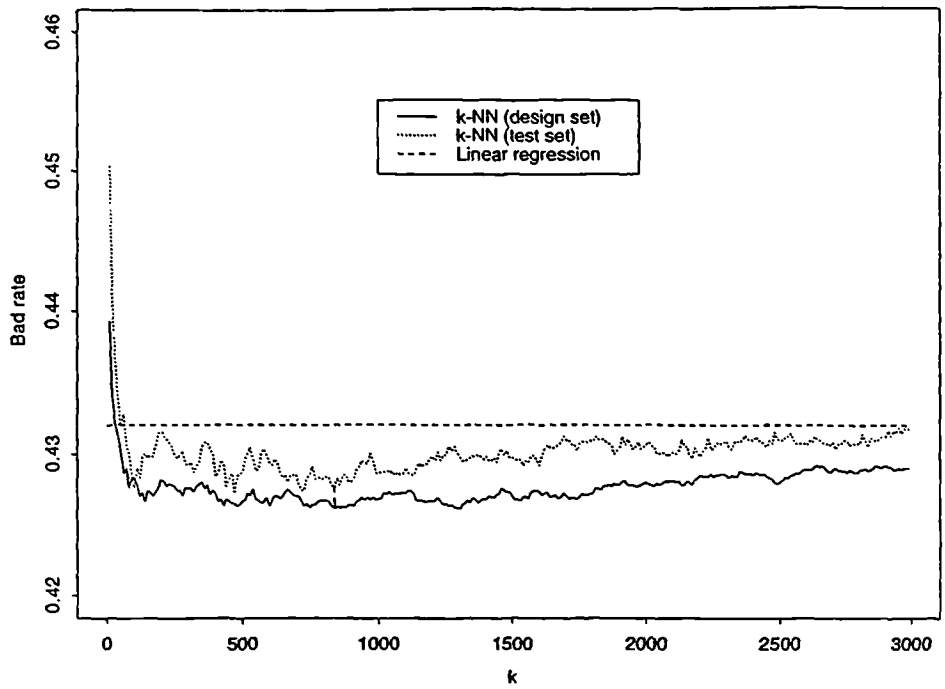


Fig 8.6.6: k-NN averaged over samples

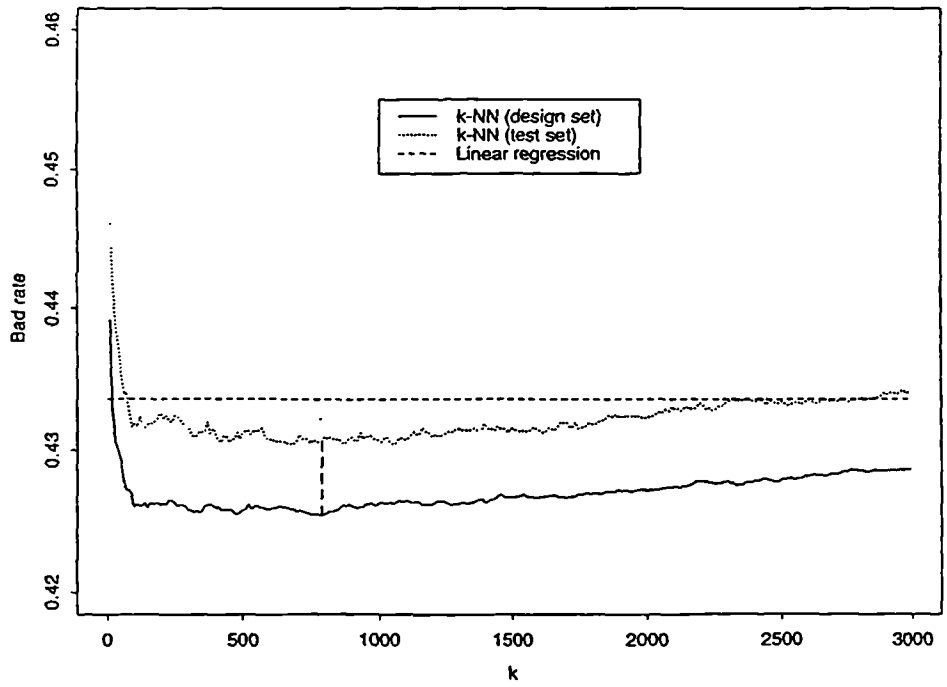
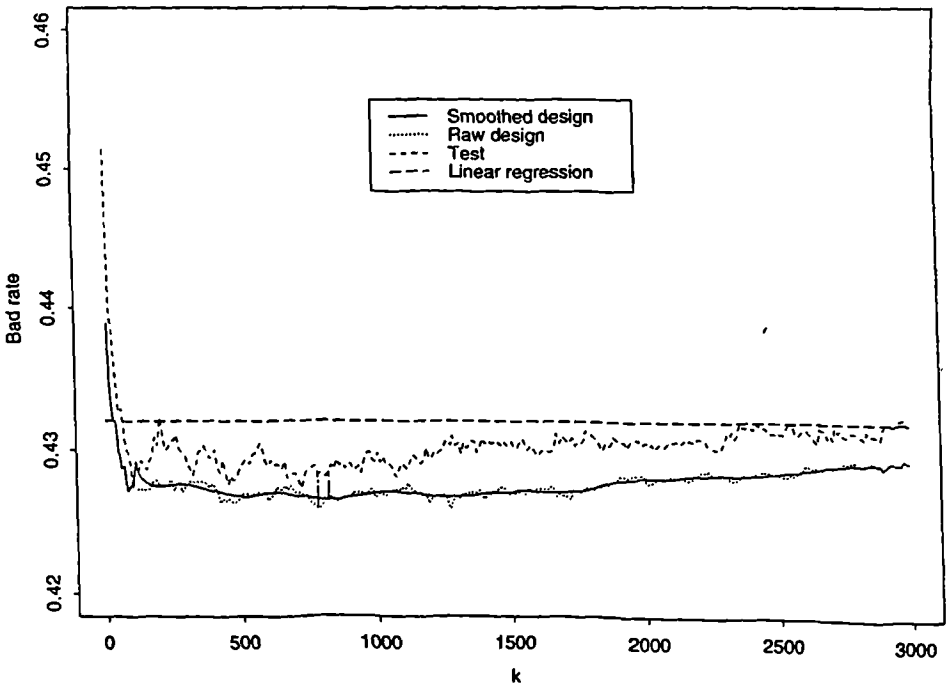


Figure 8.6.7: Smoothed design set k-NN using data set 5 and D = 1.40



8.6.1.2 Smoothing of the bad rate curves

As has already been suggested, we could simply use the values of k and D which give the lowest bad rate for the design set. However, this approach is risky because of the fluctuations in the bad rate curves for small k described in Section 8.5.1. The result that we have observed for the design set may be due to random variation and not attributable to a feature of the underlying population. In order to try to obtain a more robust choice of k we propose smoothing the bad rate curves for the design set. To find the estimated bad rate for a particular value of k we average the raw bad rates for a range of values of k around the value in question. To formalise the procedure we choose a smoothing parameter h and estimate the bad rate for $k = k_0$ by

$$s(k_0) = \begin{cases} 1/(2h+1) \sum_{i=k_0-h}^{k_0+h} r(i) & \text{for} \quad h < k_0 < 3000 - h \\ r(k_o) & \text{otherwise} \end{cases},$$

where $r(k_o)$ and $s(k_o)$ are the raw and smooth bad rates respectively for $k = k_0$.

Smoothing was not performed for very low and high values of k because we are only interested in finding estimates of the optimum values, which can be seen from Figures 8.6.1 to 8.6.6 to occur for $100 < k < 2000$. We choose to use this smoothing function because of its intuitive appeal and simplicity.

This approach to selecting k and the choice of a suitable smoothing parameter h is analogous to kernel density estimation (see for example Hand (1982)). In choosing k we wish to balance the conflicting aims of ironing out anomalies in the design set and preserving the structure of the data. For example, Figure 8.6.1 shows a sharp minimum in the design and test sets for $k = 100$ when $D = 0$. Because this k value is close to values giving much higher bad rates, the minimum is smoothed out by the above function.

Figure 8.6.7 shows an example of a smoothed bad rate curve for the design set from sample 5. The raw curves for the design and test set are added. The vertical lines represent the selected values of k using the smoothed and raw design curves.

By the same reasoning as above we could choose to use an average of the predictions from a range of values of k to predict the true class for the test sets. This is equivalent to considering smoothed versions of the test set curves. More complex weighted averages could be adopted. (This would be like working out a kernel density estimate based upon the value of the smoothing parameter selected from the design set.)

8.6.2 *k*-NN results

We now bring together the results of applying the various methods of selecting k and D discussed in the last section. For each sample we begin by selecting the value of D which gives the lowest overall bad rate in the design set (we choose to use smoothed results). Having focused on a particular value of D, we select a value (or range of values) of k from the design set (with or without smoothing) and use this value to classify the test set. The results for the test set could then be smoothed as explained in the last section.

Table 8.6.1 shows the bad rates at a 70% acceptance rate for each sample using the four combinations of smoothing/non-smoothing of the design and test set curves.

	Design (smoothing)			Design (no smoothing)			D
sample	Test (s)	Test (ns)	k	Test (s)	Test (ns)	k	
1	44.02	43.96	1260	44.01	44.01	1160	1.6
2	43.66	43.63	690	43.90	43.90	340	1.8
3	43.38	43.45	200	43.42	43.46	150	1.4
4	41.55	41.55	750	41.85	41.85	100	1.4
5	42.86	42.84	820	42.85	42.83	780	1.4

Table 8.6.1: Bad rates and selected values of k using smoothing/no smoothing of the design and test set curves for each sample.

The table shows that the choice of k is highly sensitive to the sample and whether smoothing is employed. However, since the bad rate curves are fairly flat over a wide range of k and D values, a large change in k does not give rise to much change in performance.

Although the method of smoothing employed is fairly crude it does lead to more consistent results than the case where no smoothing is used. It is clear from the results that smoothing of the design set curves is more effective than smoothing of the test set curves. In particular, for Samples 2 and 4 the results are between 0.24 to 0.3% better when the value of k is chosen from the smoothed design set results. For the other samples there is little difference in the results whether smoothing is employed or not. This indicates that if the raw curve minimum represents a real feature of the population then smoothing the bad rates will not lead to a significant change in performance; however, smoothing will help to reduce the chance of selecting a k which is only a minimum in one sample (due to random variation).

One could consider more complex smoothing functions and this might lead to improvements in the overall performance of the method. However, we chose not to do this because the shallow nature of the bad rate curves means that the performance of the method is not particularly sensitive to the choice of k. In the rest of this chapter we focus on using a smoothed version of the design set curves to select k, but do not perform smoothing on the test set curves when assessing performance.

In order to provide a baseline against which to compare our k-NN predictions Table 8.6.2 shows the bad rates for each sample from linear regression models (and from the k-NN method with k selected from the smoothed design set curve). The k-NN classifier with adjusted Euclidean metrics performs better for Samples 2-5 and linear regression performs slightly better for Sample 1. The result for Sample 1 gives grounds for being cautious in our interpretation of the results.

Sample	Linear	k-NN
1	43.87	43.96
2	44.03	43.63
3	43.54	43.45
4	42.16	41.55
5	43.20	42.84

Table 8.6.2: Bad rates for linear regression and k-NN models for samples 1-5.

The averaged bad rates for linear regression, the k-NN method with the standard Euclidean metric (i.e. $D = 0$) and the adjusted Euclidean metric are 43.36, 43.25 and 43.09 respectively. These results indicate that use of the k-NN method with adjusted Euclidean metrics can lead to improvements over use of the standard Euclidean metric and linear regression. However, applying the likelihood ratio test (Section 5.3.2), the individual sample differences are not significant. There is a need for application of our k-NN methodology to other data sets to assess whether the differences in bad rate, although small, are consistent across samples. If the difference is real then it could result in large savings for the credit grantor.

8.6.2.1 Removing the decision tree

We investigated the effect of omitting the decision tree characteristic from the set of characteristics used to design scorecards. It was not found to make a significant difference to the relative performance of the k-NN and linear regression methods for any of the samples used in this chapter. To illustrate this we describe the results for samples 1 and 4. Table 8.6.3 shows the bad rates at a 70% acceptance rate for the linear and k-NN classifiers with and without the decision tree.

	Sample 1 Decision tree		Sample 4 Decision tree	
	-	+	-	+
Linear	44.13	43.87	42.64	42.16
k-NN	44.28	43.96	41.94	41.55

Table 8.6.3 Bad rates at a 70% acceptance rate for the linear and k-NN classifiers with and without the decision tree.

For both these samples the differences between the linear and k-NN performance are similar with and without the decision trees. To further explore this Figures 6.8 and 6.9 show the bad rate curves for the k-NN method (for the test set) and linear regression with and without decision trees. These figures show that the effect of omitting the decision tree is sensitive to the value of k in the k-NN method. There do exist ranges of k values where omitting the

Fig 8.6.8: Comparison of classifiers without the decision tree characteristic (sample 4)

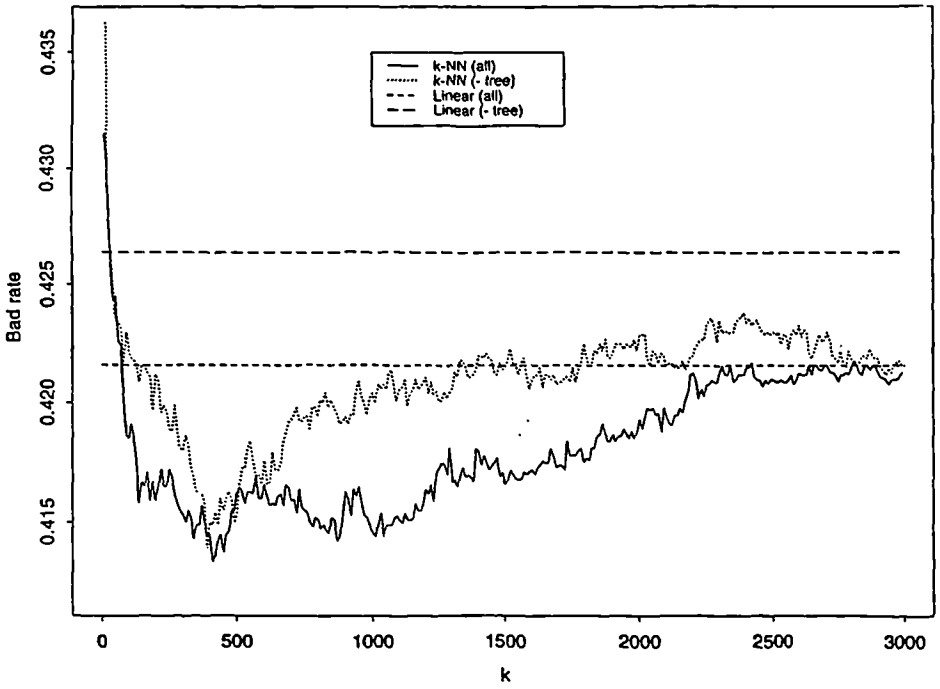
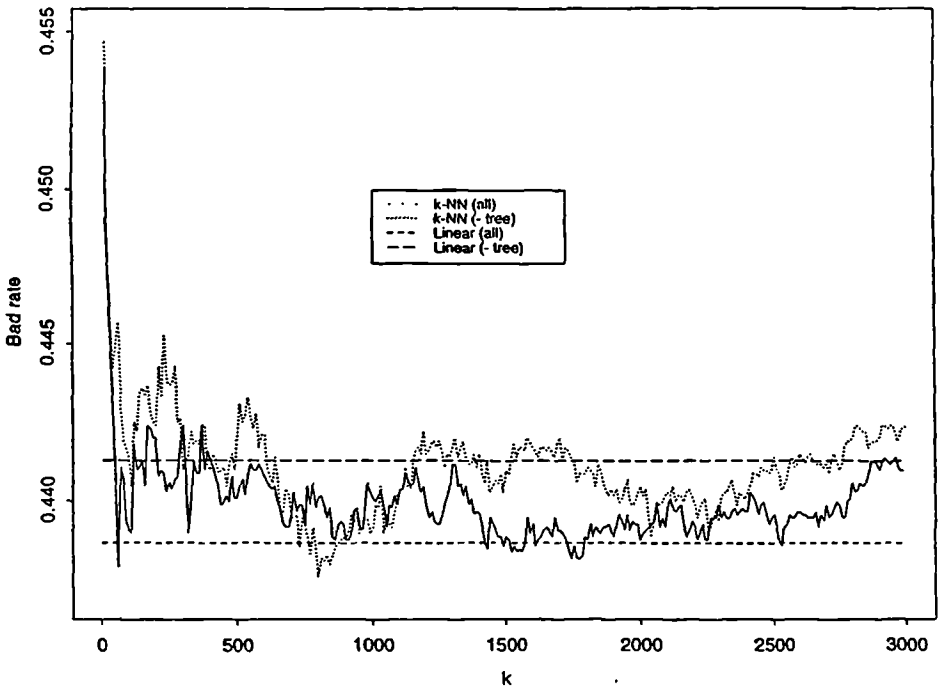


Fig 8.6.9: Comparison of classifiers without the decision tree characteristic (sample 1)



decision tree does lead to a relative improvement in the performance of the *k*-NN method.

8.6.2.2 Comparisons with other methods

In order to assess further the potential of the *k*-NN method with adjusted Euclidean metrics for credit scoring, we extended the comparisons to include logistic regression (Chapter 7), decision trees and decision graphs (as described in Section 4.9). Logistic regression and decision trees were included to represent popular parametric and non-parametric credit scoring techniques. Decision graphs were included in the comparisons to represent a recent development in classification methodology.

Table 8.6.4 shows the bad rates averaged over samples for the range of classification techniques considered. The results show that the *k*-NN method with adjusted Euclidean metrics gives the lowest bad rate.

Method	Bad rate
k-NN (any *D*)	43.09
k-NN ($D = 0$)	43.25
Logistic regression	43.30
Linear regression	43.36
Decision graphs/trees	43.77

Table 8.6.4: A summary of the averaged bad rate results using different classification techniques.

These results represent the expected performance of the different classification techniques when applied to the population from which our sample is drawn. We have proposed a practical *k*-NN classification rule which performs well, achieving the lowest overall bad rate. It was also found that the adjusted Euclidean metric led to an improvement over the standard Euclidean metric. However, if the classification rule is to be implemented, it is necessary to assess how robust it is to changes in the population over time. For this reason, in Section 8.7 we consider the use of a training sample from one population to classify applicants for credit in a future population.

8.7. An examination of the robustness of the *k*-NN method

This section explores the relative performance of the *k*-NN and linear regression methods when models are constructed using one population and applied to a different population.

8.7.1 The data

Samples from two different time periods were used in this analysis in order to simulate the real life scenario where scorecards are constructed on a sample at one point in time and then applied to applicants in future time periods. The first data set (Sample A) represents a sample collected in a previous time period for which the true creditworthiness of all applicants is known. The second data set (Sample B) represents a sample from the current population of applicants on whom an accept/reject decision needs to be taken. We use a design set from the past sample to classify the applicants in the current sample. For completeness we also classify an independent test set from the past sample.

Table 8.7.1 gives a summary of the two data sets. One immediate difference between Samples A and B is the percentage of bads in the sample (51.4/7% against 46.5%). This was due to a strategy of not accepting any of the "worst rejects" for Sample B. Note also that we restrict attention to the mini stage of the vetting procedure.

	Number of variables	Number of classes	Number of cases	% of bads in full sample
Design set (A)	21	2	14032	51.7
Test set (A)	21	2	3915	51.4
Current sample (B)	21	2	4187	46.5

Table 8.7.1: A description of Samples A and B.

Characteristics were selected for inclusion in the analysis using stepwise linear regression. We chose to use the same set of characteristics for both techniques,

despite the fact that the characteristics chosen may be sub-optimal for the k-NN method. This was to standardize comparisons and to avoid giving any unfair advantage to the k-NN method. As with previous studies in this chapter, the data was put into weights of evidence form. The assessment criterion was the bad rate at a 70% acceptance rate.

8.7.2 Classification results

The k-NN method with the adjusted Euclidean metric d_6 and the design sample A was used to classify the test sample A and the current sample B. A scorecard constructed on the design set A using linear regression was used to provide a baseline to compare the k-NN results against.

Figures 8.7.1 and 8.7.2 show the bad rate curves for the design set A, the test set A and the current sample B using $D = 0$ and $D = 1.6$ respectively. The bad rate curves for $D = 1.6$ are typical of the curves obtained using D from the optimal region ($1.5 < D < 2$). The broken lines represent the performance of the linear scorecard on the two test samples.

Figure 8.7.1 shows that when $D = 0$, the k-NN classifier does not do as well as linear regression on Sample B (the bad rate curve for the test set is above the linear regression line for all k). On the other hand, it gives superior performance for the test sample A for most k. This indicates that the k-NN method with the standard Euclidean metric is less robust to changes in the population than linear regression. The effect of using the adjusted Euclidean metric, as shown in Figure 8.7.2, is to improve the k-NN performance to about the level of the linear regression classifier on Sample B. To make this more evident, Figure 8.7.3 shows the smoothed test set curve for Sample B with $D = 1.6$. It fluctuates about the line representing the performance of linear regression. By increasing the smoothing parameter the bad rate curve converges to the linear regression line. However, this does not lead to an improvement in performance.

In order to make a more exact comparison between classifiers we considered the performance of the k-NN classifier with estimated parameters. We used the same procedure for estimating k and D that was described in Section 8.6.

Fig 8.7.1: k-NN for samples A and B with D = 0

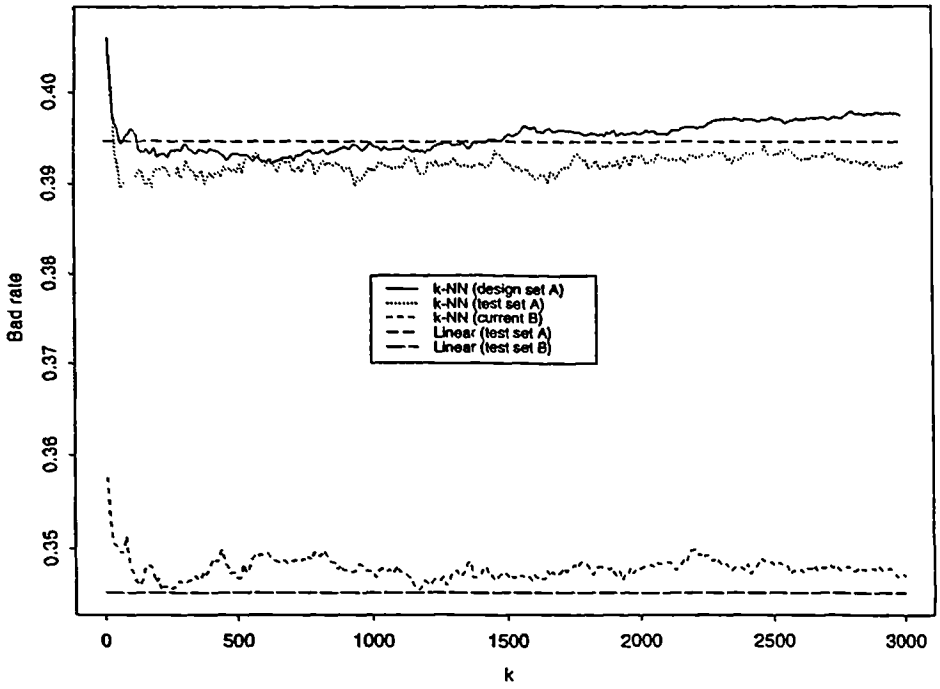


Fig 8.7.2: k-NN for samples A and B with D = 1.6

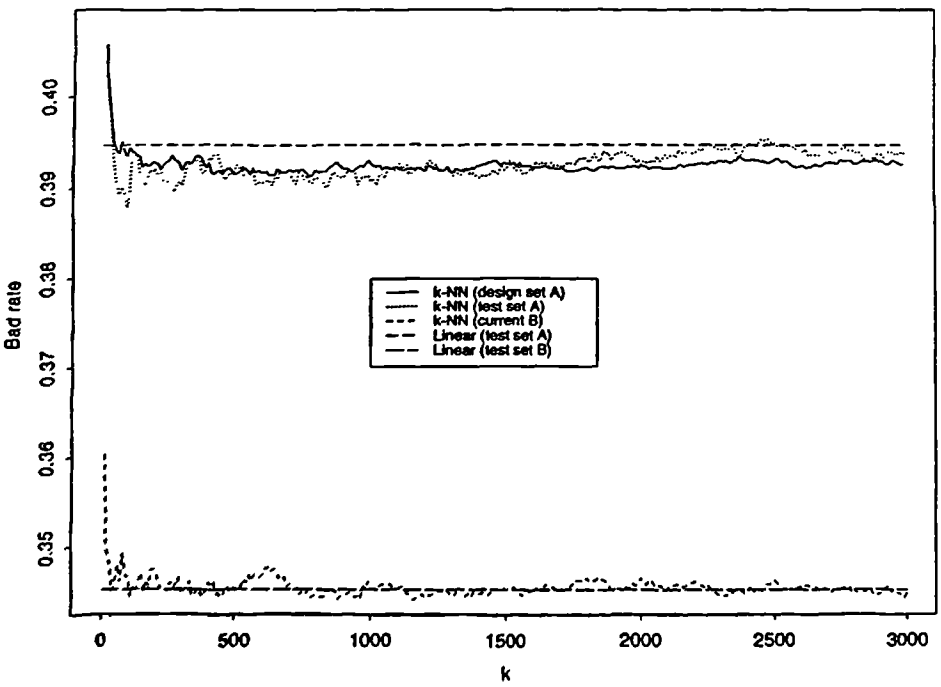


Fig 8.7.3: Smoothed bad rate curve for test sample B with D = 1.6

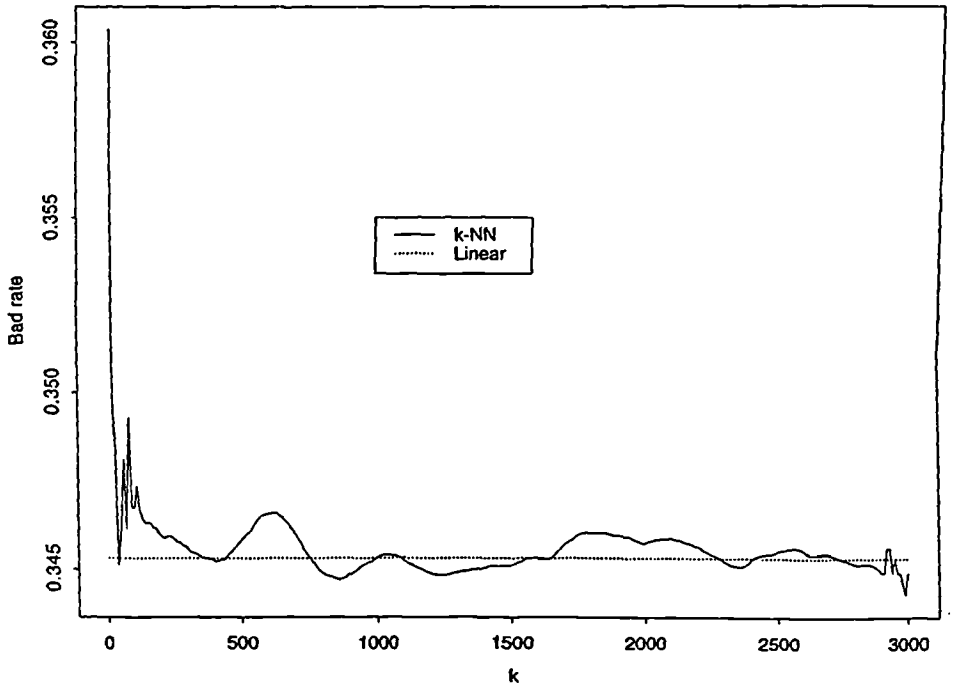


Fig 8.7.4: k-NN ranks against linear ranks

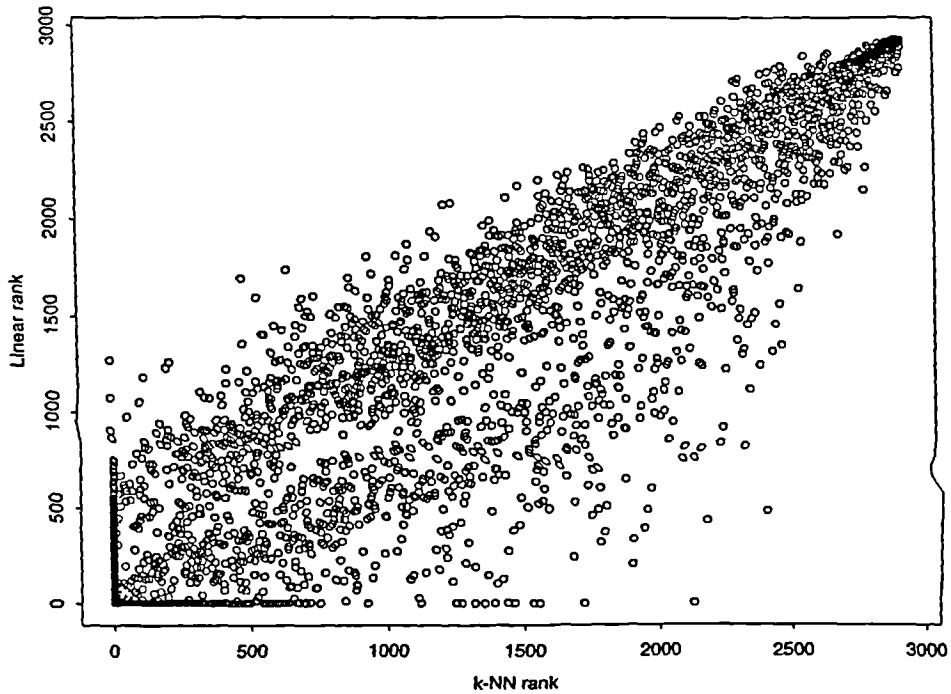


Table 8.7.2 shows the bad rates and optimal k obtained using smoothed and unsmoothed design set curves for different D.

Design (smoothing)			Design (no smoothing)			D
Sample A	Sample B	k	Sample A	Sample B	k	
39.21	34.87	640	39.27	34.88	630	0.0
39.14	34.60	720	39.17	34.67	680	1.5
39.05	34.52	720	39.12	34.76	640	1.6
39.11	34.63	690	39.18	34.53	490	1.7

Table 8.7.2: Classification results for Samples A and B using the k-NN method with smoothed/unsmoothed results.

The optimal value of D was estimated as 1.6 from the design set. Table 8.7.3 shows the classification results for linear regression and the k-NN method with optimal parameters (using the smoothed design set results to estimate k).

	Sample A	Sample B
k-NN (D = 1.6)	39.05	34.52
Linear regression	39.48	34.53

Table 8.7.3: Classification table for the k-NN method and linear regression using Samples A and B.

The results for Sample A are consistent with previous studies, giving further evidence that the k-NN method can outperform linear regression on samples from the same population. However, the two classifiers give almost identical performance for Sample B. This suggests that the k-NN methodology is less robust to changes in the nature of the population. Perhaps, this should not come as a great surprise: one advantage of a non-parametric, locally fitting model like the k-NN method is that it can identify subtle features of the population that cannot be picked up using a rigid parametric model like linear regression. When the structure of the population changes this advantage may be lost. The parametric constraints of linear regression may, in fact, be an advantage if one is trying to fit a model to a rapidly evolving population. In

this way knowledge of the underlying population gives a valuable pointer as to which type of approach to classification is most appropriate.

8.7.3 Hybrid methods

We have seen above that the k-NN method and linear regression give similar performance when applied to a sample from a future population (sample B). However, examination of swapsets reveals that the two classifiers are identifying subtly different groups of applicants. To see this we consider Table 8.7.4 which shows the numbers of goods and bads exclusively accepted under one classifier when $D = 1.6$ and $k = 500$.

	Goods	Bads
k-NN	73	122
Linear	77	118

Table 8.7.4: Swapsets from the k-NN and linear regression methods applied to sample B.

The table shows that there are very similar proportions of goods and bads in each swapset (applying Fisher's exact test the p-value is 0.755). This is why the two classifiers give similar performance. (In fact, the overall bad rates are 34.69% for the k-NN method and 34.55% for linear regression.) However, the total number of applicants in the swapsets (those accepted under one classifier and rejected under the other) is relatively high at 390 (or 13%) of the 2932 accepted applicants. It was this observation that suggested the idea of trying to combine the two methods to form a hybrid classifier.

As a first approach we ranked the $\hat{P}(g|\mathbf{x})$ for each applicant, $\mathbf{x}$, in the test set under the two classifiers. Figure 8.7.4 shows a scatter plot of the ranks for each applicant. This indicates that there is considerable variation in the ordering of the test set under the two classifiers. There appear to be two clusters of points suggesting that there are two identifiable groups of applicants which are assessed differently under the two classifiers.

Let K_i and L_i denote the ranks of the ith point in the test set under the k-NN and linear regression classifiers respectively. Then we define the hybrid rank of the ith point in the test set, H_i to be:

$$H_i = w * K_i + (1-w) * L_i, \text{ where } w \text{ is a weighting term.}$$

We note that by setting $w = 0$ the hybrid rank reduces to the linear rank and by setting $w = 1$ the hybrid rank reduces to the k-NN rank. In general one would expect to be able to improve performance over the linear and k-NN classifiers by varying the weight w. The difficulty is in choosing suitable (and robust) values for w.

In this exploratory study we chose to use the hybrid ranks, H_i, to accept 70% of the test sample and consider performance for different values of the weighting term, w. Table 8.7.5 shows the bad rates for different values of the weighting term, w.

w	0.0	0.2	0.4	0.5	0.6	0.8	1.0
Bad rate	34.55	34.38	34.31	34.31	34.24	34.28	34.69

Table 8.7.5: Bad rates from the hybrid classifier for different weights.

By taking $w = 0.6$ we are able to reduce the bad rate below the level of both individual classifiers to 34.24. The practical problem is that, as with estimating k and D, we cannot select w from the test set. Further work is needed to estimate w from the design set.

In the example considered above the k-NN and linear classifiers gave very similar performance when applied separately. By taking a hybrid classifier we were able to produce a slight improvement in performance. We now consider an example (using Sample 4 from Section 8.6) where one classifier is superior to the other. We use the k-NN method with $D = 1.4$ and $k = 750$. In this case the k-NN classifier gives a bad rate of 41.55 and the linear regression classifier gives a bad rate of 42.20 (see Table 8.6.2). Table 8.7.6 shows the resulting swapsets.

	bads	goods
Linear	79	19
k-NN	61	37

Table 8.7.6: Swapsets for data set 4 from Section 6.

Applying the significance test from Section 5.3.1, the *p*-value is 0.0025. Consequently, we reject the null hypothesis of no difference between the bad rates in the swapsets. This gives some evidence that the *k*-NN classifier is outperforming the linear classifier.

The hybrid procedure was adopted as described above. After ranking the probability estimates under the two classifiers, the hybrid rank was calculated for different w and used to accept 70% of the test sample. Table 8.7.7 shows the resulting bad rates.

w	0.0	0.2	0.4	0.6	0.8	1.0	1.17
Bad rate	42.20	41.99	41.82	41.68	41.64	41.55	41.47

Table 8.7.7: Bad rates from the hybrid classifier for different weights.

The table shows that the hybrid classifier can only just improve performance beyond the level achieved by the *k*-NN method applied separately (a bad rate of 41.47 when $w = 1.17$). This is not a sufficient improvement to be achievable in practice.

As an alternative approach to hybrid classification we investigated applying classifiers sequentially. In other words accepting a proportion p (< 0.7) of the sample under one classifier and then accepting a further proportion (0.7-p) from those rejected the first time, by applying the second classifier. This gave very similar results to the hybrid ranking approach described above.

To conclude we have not yet found sufficient evidence to justify the hybrid classifier approach. However, our analysis was limited because we only considered two examples and more extensive testing is required. It is our belief that this approach could prove to be a successful way of adding some improvement to the performance of the *k*-NN and linear regression methods.

8.8 Conclusions

In this chapter we have considered a new approach to discriminating between good and bad applicants for credit using the k-Nearest Neighbour method. We began by reviewing the standard methodology with particular attention given to the selection of a distance measure. We proposed application of the k-NN method using an original distance measure, an adjusted version of the standard Euclidean metric. An initial study was carried out to investigate properties of the classifier and bounds on performance were calculated. Plots of bad rate were considered for different values of the parameters k and D. We saw that, in fact, our k-NN classification rule is fairly insensitive to the choice of these parameters and, in particular, the curves of bad rate against k have surprisingly flat valleys. Other interesting features of the results were discussed such as the high optimal k. A comparison study showed that the adjusted Euclidean metrics perform at least as well as other standard metrics, such as the city block metric and the general Minkowski distance.

In Section 8.6 we described a more extensive and realistic study of the performance of the k-NN classifier. A practical strategy for selecting values of the parameters k and D was proposed using smoothed functions of the bad rate curves. A comparison was made between the performance of the k-NN method and a range of other classification techniques. Linear and logistic regression and decision trees were selected to represent the accepted credit scoring techniques and decision graphs were included to represent a recent development in the classification literature. It was found that the k-NN method peformed well, achieving the lowest expected bad rate. It was also found that the adjusted Euclidean metric led to an improvement over the standard Euclidean metric.

Further testing with a data set from a future population was carried out to assess the robustness of the k-NN classifier to changes in the population. It was found that linear regression was more robust, achieving similar performance on the future sample (despite doing less well on the original population). Although the linear and k-NN classifiers gave similar performance for the future sample, it

was found that they identified subtly different types of applicant. This motivated the idea of combining the two classifiers to form a hybrid classifier.

There remains one other computational limitation of the k-NN method as a practical technique for credit scoring: to assess one new applicant, one needs to calculate the distances to all the points in a design set. This is computationally a lot more expensive than techniques such as linear regression where each applicant receives a score from the sum of his/her attribute scores. This problem can be reduced by using an edited, reduced or condensed k-NN approach (see Section 8.2.5.4). Further work could investigate the influence of these approaches on the performance of the k-NN method.

Chapter 9

Conclusions

9.1 Summary of research

The aim of this thesis has been to investigate the statistical foundations necessary for building credit scoring systems. In Chapter 2 we provided an introduction to current approaches to credit scoring used by practitioners in the credit industry. This included discussion of the objectives of a credit scoring system and the nature of the data available. In Chapter 3 we reviewed previous published work on credit scoring. This was divided into literature on building credit scoring models and literature on aspects of credit granting policy. We focused on the former in this thesis. From a statistical perspective we identified three major areas requiring further research:

- the identification of suitable measures of performance and tests to compare them for different classifiers.

- the selection of appropriate analysis samples for building credit scoring models (the need for reject inference).

- the selection of appropriate classification techniques for building credit scoring models given the nature of the data and the objectives of the credit grantor.

Our conclusions are presented for each of these three areas in turn:

(1) *Assessment of performance*:
In Chapter 5 we considered different approaches to measuring the performance of credit scoring models (both relative and absolute performance). Absolute performance measures were further sub-divided into measures of discriminability and reliability. Differentiability measures assess how successful a classifier is in allocating applicants to their true class, whereas

reliability measures assess the accuracy of the predicted class membership probabilities.

The appropriate measure depends upon the commercial objectives of the lender. In our problem the lender was most interested in the bad debt amongst accepted applicants given a specific acceptance rate (with particular emphasis on 70%). Thus, the bad rate amongst the accepts was chosen as the criterion in this thesis (it is an example of a discriminability measure based upon counts of misclassifications, similar to the error rate). We discussed this criterion in some detail because it is unusual outside the credit scoring field and it imposes bounds on the levels of performance that can be achieved.

Reliability measures may be of interest to the credit grantor in some problems, such as reject inference, where accuracy is of more interest than group separation. In Section 5.2.4 attention was given to evaluating the approach to constructing reliability measures proposed by Hilden et al. (1978). A test for assessing the reliability of a classification rule was derived, given the general form of a reliability measure, and weaknesses of this approach were identified. In particular, the test is not able to identify whether a classification rule is reliable: it can only identify a subset of the cases where the rule is unreliable. This detracts from the practical value of the described approach.

The most important contribution of our work on measuring classifier performance was the proposal of two tests for assessing the relative performance of two classifiers. Both tests can be used to compare bad rates or other measures based upon the counts of misclassification, although they address subtly different questions.

The first test fixes applicants classified the same way under both classifiers and models the variation amongst applicants accepted under one and rejected under the other (using Fisher's exact test). By excluding applicants accepted under both classifiers, the test is better able to identify any small but real differences in performance.

The second test uses a likelihood based approach to compare performance. It models both possible sources of variation: first, the proportion of goods amongst applicants exclusively accepted under one of the two classifiers, and,

secondly, the proportion of applicants exclusively accepted under one classifier. It can be argued that both these sources of variation contribute to differences in classifier performance. However, this approach is also more conservative than the first test in judging a difference in performance to be significant.

(2) *Reject Inference*:
In Chapter 6 we considered different approaches to reject inference, the process of inferring the true creditworthiness of the rejects (and using this information to build improved classifiers). To begin with we carried out an experiment to evaluate which factors contribute to the bias of models built on the accepts sample. An important factor for reducing sample selection bias is the inclusion of all the characteristics used to make the original accept/reject decision. Other factors such as the adopted approach to classifier design were also discussed.

We reviewed reject inference methods proposed in the literature which attempt to use the characteristic vectors for the rejects. A likelihood based argument was used to show that the characteristic vectors for the rejects do not contain information about the parameters of the full sample model. Thus, this approach should not lead to improved performance except by:

- *Chance*. The new classifier may be better than the old one by luck

- *The use of additional information/assumptions*. Two possible approaches are the assumed distributional forms included in the mixture decomposition method, and information about $P(g|\mathbf{x})$ in the reject region in the form of a calibration sample. Several methods of reject inference that use a calibration sample were proposed. In some circumstances these methods were found to reduce the bias of extrapolation models.

- *Ad hoc adjustment of the accepts classifier*. Expert knowledge of the data may allow the adjustment of parameter estimates in a direction likely to reduce bias. For example, if a characteristic receives insufficient weighting in a scorecard built on the accepts, it might be possible to reduce bias in the classifier by multiplying the attribute scores by subjective weights.

(3) *Comparison of classification techniques*:

In Chapter 7 we made a comparison of different classification techniques for credit scoring: linear regression, logistic regression, Poisson regression, projection pursuit regression, decision trees and decision graphs. The results have shown that, given our data set:

- Linear regression is surprisingly robust to departures from the required distributional assumptions.

- There is not a statistically significant difference between the performance of the range of techniques considered. If this insensitivity of classifier performance to the technique used to build it can be confirmed for other data sets then this has implications for further research into credit scoring methodology. It indicates that research time may be more profitably spent on other aspects of the credit granting process such as identifying new sources of data.

- The results are insensitive to whether the fraud or risk definitions of credit default are used. However, future work should assess the influence of more radical innovations on classifier performance, such as use of continuous definitions of credit default.

In Chapter 8 we considered the application of the k-NN method to credit scoring using an adjusted Euclidean metric. An initial study was carried out to investigate properties of the classifier and bounds on performance were calculated. Plots of bad rate were considered for different values of the parameters k and D. We saw that, in fact, our k-NN classification rule is fairly insensitive to the choice of these parameters and, in particular, the curves of bad rate against k have surprisingly flat valleys. Other interesting features of the results were discussed such as the high optimal k. A comparison study showed that the adjusted Euclidean metrics perform at least as well as other standard metrics, such as the city block metric and the general Minkowski distance.

In Section 8.6 we described a more extensive and realistic study of the performance of the k-NN classifier. A practical strategy for selecting values of the parameters k and D was proposed using smoothed functions of the bad rate curves. A comparison was made between the performance of the k-NN method

and a range of other classification techniques. It was found that the k-NN method peformed well, achieving the lowest expected bad rate. It was also found that the adjusted Euclidean metric led to an improvement over the standard Euclidean metric.

Further testing with a data set from a future population was carried out to assess the robustness of the k-NN classifier to changes in the population. It was found that linear regression was more robust, achieving similar performance on the future sample (despite doing less well on the original population). Although the linear and k-NN classifiers gave similar performance for the future sample, it was found that they identified subtly different types of applicant. This motivated the idea of combining the two classifiers to form a hybrid classifier.

9.2 Suggestions for further research

One of the important conclusions of this thesis is that the performance of credit scoring models is fairly insensitive to the choice of classification technique. Therefore, if significant improvements are to be made, radically new approaches to modelling credit behaviour need to be identified. One possible area for future research is to consider alternative approaches to defining and modelling creditworthiness. Three possible approaches are:

- ***n*th order Markov chain models.** This approach allows creditworthiness to evolve over time with changes in state (see Section 4.11 for further details).

- **Survival analysis.** This involves modelling the time until an applicant defaults, which may be of more interest to the credit grantor than looking at performance in a constrained time period.

- **Continuous/multivariate representations of creditworthiness.** It is unrealistic to treat creditworthiness as a two state problem: in practice applicants will lie somewhere on good/bad continuum. The ideas of fuzzy set theory could be used to solve this problem.

References

Abramson, I.S. (1982) On bandwidth variation in kernel estimates-a square root law. *The Annals of Statistics* **10**, 1217-1223.

Aitchison, J. and Aitken, C.C.G. (1976) Multivariate binary discrimination by the kernel method. *Biometrika* **63**, 413-420.

Aitchison, J., Habbema, J.D.G. and Kay, J.W. (1977) A critical comparison of two methods of statistical discrimination. *Applied Statistics* **26**, 15-25.

Altman, E.I. (1968) Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, September.

Apilado, V.P., Warner, D.C. and Dauten, J.J. (1974) Evaluative techniques in consumer finance - experimental results and policy implications. *Journal of Financial and Quantitative Analysis*, March.

Avery, R.B. (1977) Credit scoring models with discriminant analysis and truncated samples. Unpublished paper.

Bazley, G. (1993) A report of the practical application of a neural network in financial decision making. Presented at *IMA conference on credit scoring and credit control III* at the University of Edinburgh, 8-10 September.

Beale, R. and Jackson, T. (1990) *Neural Computing- an Introduction.* IOP Publishing Ltd, Bristol.

Bierman, H. and Hausman, W.H. (1970) The credit granting decision. *Management Science* **16(8)**, B519-B532.

Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975) *Discrete Multivariate Analysis: Theory and Practice.* M.I.T Press, Cambridge, MA.

Blackwell, M. (1993) Measuring the effectiveness of characteristics. Presented at *IMA conference on credit scoring and credit control III* at the University of Edinburgh, 8-10 September.

Blackwell, M. and Sykes, C. (1992) The assignment of credit limits with a behaviour scoring system. *IMA Journal of Mathematics Applied in Business and Industry* **4(1)**, 1992.

Blum, M. (1974) Failing company discriminant analysis. *Journal of Accounting Research*, Spring.

Boyle, M., Crook, J.N., Hamilton, R. and Thomas, L.C. (1992) Methods for credit scoring applied to slow payers. In *Proceedings of the IMA conference on credit scoring and credit control*, ed: Thomas, L.C., Crook, J.N. and Edelman, D.B. 75-90. Clarendon Press, Oxford.

Breiman, L., Meisel, W. and Purcell, E. (1977) Variable kernel estimates of multivariate densities. *Technometrics* **19**, 135-144.

Brieman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984) Classification and Regression Trees. Wadsworth International.

Buntine, W.L. and Niblett, T. (1992) A further comparison of splitting rules for decision tree induction. *Journal of Machine Learning* **8**, 75-85.

Campbell, M.K., Donner, A., and Webster, K.M. (1991) Are ordinal models useful for classification? *Stat. Med.* **10**, 383-394.

Capon, N. (1982) Credit scoring systems: a critical analysis. *Journal of Marketing* **46** p82-91.

Carter, C. and Catlett, J. (1987) Assessing credit card applications using machine learning. In *Proceedings of the IEEE Conference on Expert Systems*, p71-79.

Chandler, G.C. and Coffman, J.Y. (1979) A comparative analysis of empirical vs judgemental credit evaluation. *Journal of retail banking* **1(2)**, 15-25.

Chandler, G.C. and Coffman, J.Y. (1983/4) Applications of performance scoring of accounts receivable management in consumer credit. *Journal of Retail Banking* **5(4)**, 1-10.

Chatterjee, S. and Barcun, S. (1970) A nonparametric approach to credit screening. *J. Am. Stat. Assoc.* **65** (329), p150-154.

Chernick, M.R., Murthy, V.K. and Nealy, C.D. (1985) Application of bootstrap and other resampling techniques: evaluation of classifier performance. *Pattern Recognition Letters* **3(3)**, 167-178.

Chow, G.K. (1957) An optimum characteristic character recognition system using decision functions. *IEEE Trans. Electronic Computers* **EC-1**, 247-254.

Corcoran, A.W. (1978) The use of exponentially-smoothed transition matrices to improve forecasting of cash flows from accounts receivable. *Management Science* **24**, 732-739.

Cover, T.M. (1968) Rates of convergence for nearest neighbour procedures. In *Proc. Hawaii Int. Conf. Syst. Sci.*, 413-415.

Cover, T.M. and Hart, P.E. (1967) Nearest neighbour pattern classification. *IEEE Trans. Inform. Theory* **IT-13**, 21-27.

Crook, J.N., Hamilton, R. and Thomas, L.C. (1992) A comparison of discriminators under alternative definitions of credit default. In *Proceedings of the IMA conference on credit scoring and credit control*, ed: Thomas, L.C., Crook, J.N. and Edelman, D.B. 217-245. Clarendon Press, Oxford.

Cyert, R.M. and Thompson, G.L. (1968) Selecting a portfolio of credit risks by Markov Chains. *J. Business* **1**, 39-46.

Cyert, R.M., Davidson, H.J. and Thompson, G.L. (1962) Estimation of the allowance for doubtful accounts by Markov chains. *Management Science* **8**, 287-303.

Davis, R.H., Edelman, D.B. and Gammerman, A.J. (1992) Macine-learning algorithms for credit-card applications. *IMA Journal of Mathematics Applied in Business and Industry* **4(1)**, 43-52.

Dawid, A.P. (1976) Properties of diagnostic data distributions. *Biometrics* **32**, 647-658.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977) Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion), *J. Roy. Statist. Soc.* **B39**, 1-38.

Dirickx, Y.M.I. and Wakeman, L. (1975) An extension of the Bierman-Hausman model for credit granting. *Management Science* **22(11)**, 1229-1237.

Dobson, A.J. (1983) *An Introduction to Statistical Modelling*. Chapman and Hall, New York.

Doreen, D.D. and Farhoomand, F. (1983) A decision model for small business loans. *Journal of Small Business* (Canada), Fall.

Draper, N.R., and Smith H. (1981) Applied regression analysis, New York: John Wiley, 2nd ed.

Dudani, S.A. (1976) The distance weighted *k*-nearest neighbour rule. *IEEE Transactions on Systems, Man and Cybernetics* **SMC-6**, 325-327.

Durand, D. (1941) Risk elements in consumer instalment financing. Financial Research Program, Study 8, National Bureau of Economic Research.

Edelman, D.B. (1992) An application of cluster analysis in credit control. *IMA Journal of Mathematics Applied in Business and Industry* **4(1)**, 81-88.

Edmister, R.O. (1972) An empirical test of financial ratio analysis for small business failure prediction. *Journal of Financial and Quantitative Analysis*, March.

Edmister, R.O. and Schlarbaum, G.G. (1974) Credit policy in lending institutions. *Journal of Financial and Quantitative Analysis* **10**, 335-356.

Efron, B. (1982) The jackknife, the bootstrap, and other resampling plans. Society for Industrial and Applied Mathematics, Philadelphia, PA.

Efron, B. (1983) Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association* **78**, 316-331.

Eisenbeis, R.A. (1977) Pitfalls in the application of discriminant analysis in business, finance and economics. *Journal of Finance* **32(3)**, 875-900.

Eisenbeis, R.A. (1978) Problems in applying discrminant analysis in credit scoring models. *J. Banking and Finance* **2**, 205-219.

Enas, G.C. and Choi, S.C. (1986) Choice of the smoothing parameter and efficiency of *k*-nearest neighbour classification. *Comp. and Maths. with Appl.* **12A(2)**, 235-244.

Epanechnikov, V.K. (1969) Non-parametric estimation of a multivariate probability density. *Theory Probab. Appl.* **14**, 153-158.

Fielding, A. (1977) Latent structure analysis. In *Exploring Data Structures*. ed: O'Muircheartaigh, C.A. and Payne, C. New York: Wiley, p125-157.

Fienberg, S.E. (1977) *The Analysis of Cross-Classified Categorical Data*, M.I.T. Press, Cambridge, MA.

Finney, P.J. (1952) *Probit Analysis*. Cambridge, MA., Cambridge University Press.

Fisher, R.A. (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7**, 179-188.

Fix, E. and Hodges, J. (1952) Discriminatory analysis, nonparametric discrimination: consistency properties. Report no. 4, project no. 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas.

Fogarty, T.C. and Ireson, N.S. (1994) Evolving Bayesian classifiers for credit control - a comparison with other machine learning methods. *IMA Journal of Mathematics Applied in Business and Industry* **5 (1)**, 63-75.

Friedman, J.H. (1979) A tree-structured approach to nonparametric multiple regression. In *Smoothing Techniques for Curve Estimation*, ed: Gasser, Th. and Rosenblatt, M., New York: Springer Verlag, p5-22.

Friedman, J.H. (1989) Regularized discriminant analysis. *Journal of the American Statistical Association*, **84**, 165-175.

Friedman, J.H., and Stuetzle, W. (1981) Projection pursuit regression. *Journal of the American Statistical Association* **76(376)**, 817-823.

Frydman, H., Kallberg, J.G. and Kao, Duen-Li. (1985) Testing the adequacy of Markov chain and mover-stayer models as representations of credit behaviour. *Operations Research* **33(6)**, 1203-1214.

Fukanaga, K., and Hostetler, L.D. (1973) Optimization of k-nearest neighbour density estimates. *IEEE Trans. Inform. Theory* **IT-19**, 320-326.

Fukanaga, K. and Flick, T.E. (1984) An optimal global nearest neighbour metric, *IEEE Trans. Pattern Anal. Mach. Intell.* PAMI-**1**, 25-37.

Fukunaga, K. and Hummels, D.M. (1987) Bias of nearest neighbour error estimates. *IEEE Transactions on pattern analysis and machine intelligence*, PAMI-**9(1)**, January.

Gates, G.W. (1972) The reduced nearest neighbour rule. *IEEE transactions on Information theory* **IT-18**, 431.

Gilbert, L.R., Menon, K. and Schwartz, K.B. (1990) Predicting bankruptcy for firms in financial distress. *J. Business Finance Account.* **17**(1), 161-171.

Goldberg, D.E. (1989) *Genetic algorithms in search, optimisation and machine learning*. Reading, MA, Addison-Wesley.

Goldstein, M. and Dillon, W.R. (1978) *Discrete Discriminant Analysis*. New York: Wiley.

Grablowsky, B.J. and Talley, W.K. (1981) Probit and discriminant functions for classifying credit applicants: a comparison. *J. Econ and Business* **33**, 254-261.

Greer, C.C. (1967) The optimal credit acceptance scheme. *Journal of Financial and Quantitative Analysis* **3**, 399-415.

Habbema, J.D.F., Hilden, J. and Bjerregaard, B. (1978) The measurement of performance in probabilistic diagnosis I. The problem, descriptive tools and measures based on classification matrices. *Methods Inf. Med.* **17**, 217-226.

Hand, D.J. (1981) *Discrimination and Classification*. New York: John Wiley.

Hand, D.J. (1982) *Kernel discriminant analysis*. Letchworth: Research Studies Press.

Hand, D.J. (1986) Recent advances in error rate estimation. *Pattern Recognition Letters* **4**, 335-346.

Hand, D.J. (1992) Statistical methods in diagnosis. *Statistical Methods in Medical Research* **1**, 49-67.

Hand, D.J. (1994) Assessing classification rules. *Journal of Applied Statistics* **21(3)**, 3-16

Hand, D.J. and Batchelor, B.G. (1978) An edited nearest neighbour rule. *Information Sciences* **14**, 171-180.

Hand, D.J. and Henley, W.E. (1993/4) Can reject inference ever work? *IMA Journal of Mathematics Applied in Business and Industry* **5**, 45-55.

Hand, D.J. and Henley, W.E. (1994) Inference about rejected cases in discriminant analysis. In *New Approaches in Classification and Data Analysis*, ed. Diday, E., Lechevallier, Y., Schader, M., Bertrand, P., and Burtschy, B. Published by Springer-Verlag, Berlin, p292-299.

Hart, P.E. (1968) The condensed nearest neighbour rule. *IEEE Transactions on Information Theory* **IT-14**, 515-516.

Henley, W.E. and Hand, D.J. (1994) A *k*-NN classifier for assessing consumer credit risk. Submitted to *The Statistician*.

Henley, W.E. and Hand, D.J. (1995) Some developments in statistical credit scoring. To appear in *Machine Learning and Statistics: The Interface*, ed. Taylor, C.C. Published by John Wiley and Sons.

Hilden, J. (1984) Statistical diagnosis based on conditional independence does not require it. *Computers in Biology and Medicine*, **14**, 429-435.

Hilden, J., Habbema, J.D.F. and Bjerregaard, B. (1978a) The measurement of performance in probabilistic diagnosis II. Trustworthiness of the exact values of the diagnostic probabilities. *Methods of Information in Medicine* **17**, 227-237.

Hilden, J., Habbema, J.D.F. and Bjerregaard, B. (1978b) The measurement of performance in probabilistic diagnosis III. Methods based on continuous functions of the diagnostic probabilities. *Methods of Information in Medicine* **17**, 238-246.

Hills, M. (1967) Discrimination and allocation with discrete data. *Applied Statistics* **16**, 237-250.

Holland, J.H. (1975) *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control and AI.* Ann Arbor, MI, The University of Michigan Press.

Hsia, D.C. (1978) Credit scoring and the equal credit opportunity act. *The Hastings Law Journal* **30**, 371-448

Joanes, D.N. (1993/4) Reject inference applied to logistic regression for credit scoring. *IMA Journal of Mathematics Applied in Business and Industry*, **5(1)**, 35-43.

Jozwick, A. (1983) A learning scheme for a fuzzy *k*-NN rule. *Pattern Recognition Letters* **1**, 287-289.

Keller, J.M., Gray, M.R. and Givens Jr., J.A. (1985) A fuzzy *k*-nearest neighbour algorithm. *IEEE Transactions on Systems, Man and Cybernetics* **SMC(4)**, 580-585.

Kendall, M. and Stuart, A. (1977) *The Advanced Theory of Statistics*, Vol.1, Fourth Edition. Published by Griffin, London.

Khoylou, J. and Stirling, M. (1993) Credit scoring and neural networks. Presented at the *IMA conference on credit scoring and credit control III* at the University of Edinburgh, 8-10 September.

Koplowitz, J. and Brown, T.A. (1981) On the relation of performance to editing in nearest neighbour rules. *Pattern Recognition* **15(3)**, 251-255.

Krzanowski, W.J. (1975) Discrimination and classification using both binary and continuous variables. *J. Am. Stat. Assoc.* **70**, 782-790.

Leonard, K.J. (1988) *Credit scoring via linear logistic models with random parameters.* Ph.D. Dissertation, Department of Decision Sciences and Management Information Systems, Concordia University, Montreal, Canada.

Leonard, K.J. (1993) Empirical Bayes analysis of the commercial loan evaluation process *Statistics and Probability Letters* **18**, 289-296.

Leonard, K.J. (1993/4) A fraud-alert model for credit cards during the authorization process. *IMA Journal of Mathematics Applied in Business and Industry*, **5(1)**, 57-62.

Little, R.J.A. and Rubin, D.B. (1987) *Statistical Analysis with Missing Data.* New York: John Wiley.

Loftsgaarden, D.O. and Quesenberry, C.P. (1965) A nonparametric estimate of a multivariate density function. *Ann. Math. Statist.* **36**, 1049-1051.

Long, M.S. (1976) Credit screening system selection. *Journal of Financial and Quantitative Analysis*, June.

Lundy, M. (1992) Cluster analysis in credit scoring. In *Proceedings of the IMA conference on credit scoring and credit control*, ed: Thomas, L.C., Crook, J.N. and Edelman, D.B. 91-107. Clarendon Press, Oxford.

McClelland, J.L. and Rumelhart, D.E. (1986) *Parallel Distributed Processing, Volume 1*. MIT Bradford Press.

McCullagh, P. (1980) Regression models for ordinal data. *Journal of the Royal Statistical Society (Series B)* **42**, 109-142.

McCullagh, P. and Nelder, J.A. (1983) *Generalized Linear Models*. Chapman and Hall, London.

McLachlan, G.J. (1975) Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. *Journal of the American Statistical Association* **70**, 365-369.

Malkovich, J.F. and Afifi, A.A. (1973) On tests for multivariate normality. *Journal of the American Statistical Association*, **68(341)**, March.

Marks, S. and Dunn, O.J. (1974) Discriminant functions when covariance matrices are unequal. *J. Am. Stat. Assoc.* **69**, no.346, June.

Mehta, D. (1968) The formulation of credit policy models. *Management Science* **15(2)**, B30-B50.

Mehta, D. (1970) Optimal credit policy selection: a dynamic approach. *Journal of Financial and Quantitative Analysis*, Dec.

Myers, J.H. and Cordner, W. (1957) Increase credit operation profits. *The Credit World*, Februrary, p12-13.

Myers, J.H. and Forgy, E.Q. (1963) The development of numerical credit evaluation systems. *J. Am. Stat. Assoc.* **58**, 799-806.

Myles, J. (1991) *The use of k-Nearest Neighbour methods in statistical pattern recognition*. Ph.D. Dissertation, Department of Statistics, The Open University.

Oliver, J.J. (1992) Decision graphs - an extension of decision trees. Technical report no: 92/173, Computer Science Department, Monash University, Vic 3168, Australia.

Oliver, J.J. and Hand, D.J. (1994) Fanned decision trees. Technical Report no: 94-5, Department of Statistics, The Open University, Walton Hall, Milton Keynes.

Oliver, R.M. (1992) The economic value of score-splitting accept-reject policies. *IMA Journal of Mathematics Applied in Business and Industry* **4(1)**, 35-42.

Orgler, Y.E. (1970) A credit scoring model for commercial loans. *Journal of money, credit and banking*, November.

Orgler, Y.E. (1971) Evaluation of bank consumer loans with credit scoring models. *Journal of Bank Research*, Spring.

Overstreet, G.A. and Kemp, R.S. (1986) Managerial control in credit scoring systems. *Journal of Retail Banking* **8(182)**, 79-86.

Parzen, E. (1962) On the estimation of probability density function and mode. *Ann. Math. Statist.* **33**, 1065-1076.

Pinches, G. and Mingo, K. (1973) A multivariate analysis of industrial bond ratings. *Journal of Finance*, March.

Reichert, A.K., Cho, C.C. and Wagner, G.M. (1983) An examination of the conceptual issues involved in developing credit scoring models. *J. Bus. and Economic Statistics* **1**, 101-114.

Rosenblatt, M. (1956) On some nonparametric estimates of a density function. *Ann. Math. Statist.* **27**, 832-837.

Shapiro, A.R. (1977) The evaluation of clinical predictions. *The New England Journal of Medicine* **296(26)**, 1509-1514.

Short, R.D. and Fukunaga, K. (1982) The optimal distance measure for nearest neighbour classification. *IEEE Trans. Inform. Theory* IT-**27**, 622-627.

Smith, P.F. (1964) Measuring risks on consumer installment credit. *Management Science*, Nov.

South, M.C., Wetherill, G.B. and Tham, M.T. (1993) Hitch-hiker's guide to genetic algorithms. *Journal of Applied Statistics* **20(1)**, 153-175.

Srinivasan, V. and Kim, Y.H. (1987) Credit granting: a comparative analysis of classification procedures. *J. Finance* **92**, 665-681.

Steenackers, A. and Goovaerts, M.J. (1989) A credit scoring model for personal loans. *Insur. Maths. Econom* **8**.

Stone, C.J. (1977) Nonparametric regression and its applications (with discussion). *Annals of Statistics* **5**, 595-645.

Terrell, G.R. and Scott, D.W. (1992) Variable kernel density estimation. *The Annals of Statistics* **20** (3), 1236-1265.

Titterington, D.M. (1992) Discriminant analysis and related topics. In *Proceedings of the IMA conference on credit scoring and credit control*, ed: Thomas, L.C., Crook, J.N. and Edelman, D.B. 53-73. Clarendon Press, Oxford.

Titterington, D.M., Murray, G.D., Murray, L.S., Spiegelhalter, D.J., Skene, A.M., Habbema, J.D.F. and Gelpke, G.J. (1981) Comparison of discrimination techniques applied to a complex data set of head injured patients (with discussion). *Journal of The Royal Statistical Society (Series A)* **144**, 145-175.

Todeschini, R. (1989) *k*-nearest neighbour methods: the influence of data transformations and metrics. *Chemometrics Intell. Labor. Syst.* **6**, 213-220.

Toussaint, G.T. (1974) Bibliography on estimation of misclassification. *IEEE Transactions on Information Theory* **20**, 472-479.

Tukey, P.A. and Tukey, J.W. (1981) Data-driven view selection: Agglomeration and sharpening. In *Interpreting Multivariate Data*, ed. Barnett, V., published by John Wiley, Chichester.

Upton, G.J.G. (1982) A comparison of alternative tests for the 2 X 2 comparative trial. *Journal of the Royal Statistical Society (Series A)*, **145**, 86-105.

van Kuelen, J.A., Spronk, J. and Corcoran, A.W. (1981) Note on the Cyert-Davidson-Thompson doubtful accounts model. *Management Science* **27**, 108-112.

Wallace, C.S. and Boulton, D.M. (1968) An information measure for classification. *Computer Journal* **11**, 185-194.

Wallace, C.S. and Patrick, J.D. (1993) Coding decision trees. *Machine Learning* **11**, 7-22.

Wiginton, J.C. (1980) A note on the comparison of logit and discriminant models of consumer credit behaviour. *J. Fin. Quantitative Analysis* **15**, 757-770.

Wilkie A.D. (1992) Measures for comparing scoring systems. In Proceedings of the IMA conference on credit scoring and credit control, ed: Thomas, L.C., Crook, J.N. and Edelman, D.B. 123-138. Clarendon Press, Oxford

Yates, F. (1984) Tests of significance for 2 X 2 contingency tables. Journal of the Royal Statistical Society (Series A) **147(3)**, 426-463.

Yoon, Y., Swales JR, G. and Margavio, T.M. (1993) A comparison of discriminant analysis versus artificial neural networks. *J. Opl Res. Soc.* **44(1)**, 51-60.

Zadeh, L.A. (1965) Fuzzy sets. *Inf. Control.* **8**, 338-353.